Journal of Political Economy

EDITED BY JACOB A. FRENKEL, ROBERT E. LUCAS, JR.,
SAM PELTZMAN, AND GEORGE J. STIGLER
IN CO-OPERATION WITH OTHER MEMBERS OF THE
DEPARTMENT OF ECONOMICS AND THE
GRADUATE SCHOOL OF BUSINESS OF
THE UNIVERSITY OF CHICAGO
AND OUTSIDE REFEREES

Volume 89

February–December 1981

THE UNIVERSITY OF CHICAGO PRESS
CHICAGO, ILLINOIS

PUBLISHED FEBRUARY,
APRIL, JUNE, AUGUST, OCTOBER,
DECEMBER 1981
BY THE UNIVERSITY OF CHICAGO PRESS

Journal of Political Economy

Volume 89, Number 1, February 1981

The University of Chicago Press

JOURNAL OF
POLITICAL ECONOMY

Edited by

JACOB A. FRENKEL SAM PELTZMAN
ROBERT E. LUCAS, JR. GEORGE J. STIGLER

In cooperation with OTHER MEMBERS of the DEPARTMENT OF ECONOMICS and the GRADUATE SCHOOL OF BUSINESS of the UNIVERSITY OF CHICAGO AND OUTSIDE REFEREES

Editorial Assistants: VICKY M. LONGAWA and LISE A. PLOTKIN

The Journal of Political Economy (ISSN 0022-3808) is published bimonthly in February, April, June, August, October, and December by the University of Chicago Press. Subscription rates, U.S.A.: institutions, 1 year $30.00, 2 years $54.00, 3 years $76.50; individuals, 1 year $22.00. 2 years $39.60, 3 years $56.10. Student subscription rate, U.S.A.: 1 year $16.00 (letter from professor must accompany subscription). Other countries add $2.50 for each year's subscription to cover postage. Single copy rates: institutions $5.00, individuals $4.00. Back issues are available from 1962 (vol. 70). Make all remittances payable to *Journal of Political Economy*, The University of Chicago Press, in United States currency or its equivalent. **Business correspondence** should be addressed to The University of Chicago Press, 5801 Ellis Avenue, Chicago, Illinois 60637.

Claims for missing numbers should be made within the month following the regular month of publication. The publishers expect to supply missing numbers free only when losses have been sustained in transit and when the reserve stock will permit.

Letters to the editors and manuscripts should be addressed to the Editor of the *Journal of Political Economy*, 1126 East 59th Street, Chicago, Illinois 60637. **Manuscripts should be submitted in triplicate, accompanied by a $40.00 submission fee made payable to the Journal.** The proceeds from the submission fees are used to pay for refereeing services. Accepted manuscripts must be typed according to the University of Chicago *Manual of Style*. References should be typed double-spaced at the end of the article. Footnotes should be numbered in sequence and double-spaced following the references. Tables should follow the footnotes. Originals of the figures, drawn in india ink, should be submitted if the manuscript is accepted. Abstracts not exceeding 100 words should be submitted in duplicate along with the manuscript.

Reprinted volumes 1–72 available from Walter J. Johnson, Inc., 355 Chestnut Street, Norwood, New Jersey 07648. Volumes available **in microfilm** from University Microfilms, 300 North Zeeb Road, Ann Arbor, Michigan 48106; **in microfiche** from Johnson Associates, P.O. Box 1017, Greenwich, Connecticut 06830.

Notice to subscribers: If you change your address, please notify us and your local postmaster immediately, giving *both* your old and your new address. *Allow four weeks for the change.* **Postmaster:** Send address changes to *Journal of Political Economy*, 5801 Ellis Avenue, Chicago, Illinois 60637.

Second-class postage paid at Chicago, Illinois, and at additional mailing office.

Journal of Political Economy

Volume 89 Number 1 February 1981

Articles

Comments

Positive Time Preference

Mancur Olson and Martin J. Bailey

University of Maryland

The case for positive time preference is absolutely compelling, unless there is an infinite time horizon with the expectation of unending technological advance combined with what we call "drastically diminishing marginal utility." This finding holds both in the positive and normative senses. A corollary is that savings are interest elastic.

Most writers on capital theory have taken it for granted that most people systematically prefer present to future consumption. This positive rate of time preference is usually ascribed to a "faulty telescopic faculty," to "myopia," or to "impatience," even though this implies an irrationality or continuing perceptual bias that is inconsistent with the behavioral assumptions economists usually use. To our knowledge no one has ever provided convincing evidence that there is in fact normally positive time preference, or even specified an empirical test capable of determining whether there is or not. Apparently the belief in myopia often rests on subjective judgments about the relative utility levels of other individuals across different time periods that have very limited scientific value. In other cases a preference for present over future consumption is erroneously supported with evidence that, as we shall argue, need reflect nothing more than a rational adaptation to positive interest rates generated by the fact that capital goods can be useful in production. In still other instances time preference is supposed to follow directly from the fact that there is

We are indebted to Gary Becker, Christopher Clague, Donald Dewey, Robert Dorfman, Tjalling Koopmans, Peter Murrell, George Stigler, Paul Wonnacott, and Howell Zee for helpful comments and suggestions. Part of the work on this article was supported by grants from the Environmental Protection Agency, the National Science Foundation, and Resources for the Future.

[*Journal of Political Economy,* 1981, vol. 89, no. 1]

always uncertainty about the future, so the consumer cannot be sure whether any assets that he acquires by saving will retain their value or even that he will be alive to enjoy any fruits of his saving. In fact, uncertainty about future economic conditions and the duration of one's life can increase as well as decrease the incentive to save.

This paper develops an approach to intertemporal choice that makes it possible to determine from actual household choices in certain circumstances whether or not there is positive time preference. The empirical tests on ordinary saving and consumption behavior that grow out of this approach suggest that a decidedly positive rate of time preference is typical if not almost universal. There is an alternative interpretation of the evidence that permits doubts in cases where the planning horizon is infinite, and this we label "drastically diminishing marginal utility." This intriguing phenomenon could in principle be widespread. In long-run contexts it is so similar to time preference in its implications that the two are difficult to distinguish. Nonetheless, especially when short- as well as long-run contexts are considered, the case that there is often time preference becomes compelling, even if drastically diminishing marginal utility should also be present.

We shall also show that those who advocate certain environmental and resource policies on the ground that the interests of all future years and generations ought to be weighted equally with our concern for utility in the present are contradicted by their own behavior. We will similarly show that it is logically inconsistent to argue that there is positive time preference and that household savings do not generally respond to changes in interest rates.

I. Positive Time Preference and Diminishing Marginal Utility

Often the idea of time preference appears in an impressionistic form without any workable definitions (see, e.g., Pigou 1960; Jevons 1965), though there are also a number of quite useful efforts to obtain a precise understanding of the matter.[1] Interestingly enough, the clearest conception of positive time preference that we have been able to find was in Böhm-Bawerk's original account:

> *Present goods have in general greater subjective value than future (and intermediate) goods of equal quantity and quality. . . .*
>
> A *first* principal cause capable of producing a difference in

[1] See the helpful discussion in Marglin (1963), Feldstein (1964), and Hirshleifer (1970). There are also important articles that appeared somewhat earlier, notably Samuelson (1937), Strotz (1956), Debreu (1959), and Koopmans (1960).

> value between present and future goods is . . . if a person suffers in the present from appreciable lack of certain goods, or of goods in general, but has reason to hope to be more generously provided for at a future time, then that person will always place a higher value on a given quantity of immediately available goods than on the same quantity of future goods. This situation occurs with very great frequency in our economic life. . . .
>
> We must now consider a *second* phenomenon of human experience—one that is heavily fraught with consequence. That is the fact that we feel less concerned about future sensations of joy and sorrow simply because they do lie in the future, and the lessening of our concern is in proportion to the remoteness of that future. Consequently we accord to goods which are intended to serve future ends a value which falls short of the true intensity of their future marginal utility. [1959, pp. 265–73]

This quotation suggests two distinct elements, each of which might influence intertemporal choice: (1) diminishing marginal utility of consumption at any given time, where utility is an unchanging function of the amount of consumption at the time, and (2) discounting of future versus present utility. This distinction is so fundamental that we believe there can be no adequate explanation of intertemporal choice without it, yet it is too often ignored. Each of these two elements, moreover, naturally suggests a testable hypothesis that can be proven false. Bailey, Olson, and Wonnacott have shown elsewhere that "The Marginal Utility of Income Does Not Increase" (1980) and in at least a great majority of cases is actually decreasing. In this paper we accordingly accept Böhm-Bawerk's assumption that there is diminishing marginal utility to consumption in each period and go on to explore whether the available evidence is consistent with the hypothesis that utility in the future is valued less than utility in the present. In order to do this, and to obtain the full range of hypotheses that can be drawn with Böhm-Bawerk's distinction, we shall first develop a more formal statement of the distinction.

Following a suggestion by Hirshleifer (1970),[2] we represent utility at moment t by a function $v(C_t)$; the consumption C_t is time subscripted, but the function is time invariant, depending only on the level or quantity of C_t, regardless of t. Then with time measured in terms of discrete periods of equal length, we write the consumer's

[2] The Hirshleifer analysis in turn has precedents in the articles by Samuelson (1937), Strotz (1956), Debreu (1959), and Koopmans (1960) cited in the preceding footnote.

utility function to be maximized at time zero, subject to a wealth budget constraint, as

$$U = \sum_{t=0}^{T} \frac{v(C_t)}{(1+\eta)^t}, \tag{1}$$

where the consumer's time horizon for planning is T, which need not be finite, and where η is a constant, the rate per period of positive time preference. We assume that the first derivative v' is positive and (because of the finding mentioned before) that the second is negative, $v'' < 0$.

Now consider a household choosing among alternative combinations (C_0, C_1) of consumption at $t = 0$ and $t = 1$, holding its consumption pattern for all later periods fixed. If we take the differential of (1), setting $dC_t = 0$ for all $t \geqslant 2$, we have

$$dU = v'(C_0)dC_0 + \frac{v'(C_1)dC_1}{1+\eta}, \tag{2}$$

and setting $dU = 0$ to represent holding the household indifferent,

$$v'(C_0)dC_0 + \frac{v'(C_1)dC_1}{1+\eta} = 0,$$

or

$$\left(\frac{dC_1}{dC_0}\right)_{U = \text{const.}} = -\frac{v'(C_0)}{v'(C_1)}(1+\eta). \tag{3}$$

The household will demand an increase in C_1 greater than the reduction in C_0, that is,

$$\left|\frac{dC_1}{dC_0}\right| > 1,$$

to hold utility constant, if either of two things is true:

$$C_1 > C_0 \text{ with } \eta = 0, \tag{4}$$

or

$$\eta > 0 \qquad \text{with } C_1 = C_0. \tag{5}$$

Conditions (4) and (5) succinctly state the two causes of interest in the quoted passage of Böhm-Bawerk.

Inequality (4) follows because $v'' < 0$. Diminishing marginal utility of income implies that, if all else is equal, a household with greater consumption potential in the later period will be willing to borrow at a positive interest rate in order to make consumption in the two periods more nearly equal. This is the "first cause" of the preference

for present over future consumption mentioned in the quotation from Böhm-Bawerk.

The hypothesized second cause is represented by the constant η in the intertemporal utility function (1). When it is positive rather than zero as in (5), the household would be willing to pay some positive interest rate to increase consumption in the initial period at the expense of consumption in the later period, even with equal endowments in the two periods.

By tracing out the combinations of C_0 and C_1 to which the household is indifferent, using (1) or (3), assuming $v(C_t)$ and η are known, one can obtain an indifference curve between present and next-period consumption. That is, equation (1) is an implicit function of C_0 and C_1, when all other C_t are held constant and when U is also held constant; this function,

$$U_0 = v(C_0) + \frac{v(C_1)}{1 + \eta} + \text{const.},$$

is the equation of an indifference curve in (C_0, C_1) space. Such an indifference curve U_0 is shown in figure 1. Along the 45° line $C_0 = C_1$, and on this line any tangent to an indifference curve, such as ABD, has a slope of $-(1 + \eta)$. This can be seen by setting $C_0 = C_1$ in equation (3), or by noting that when consumption is the same in both periods, any preference for present over future consumption must be due to η.

Though the marginal rate of substitution between consumption in any two periods for a household in equilibrium is frequently called the rate of "time preference," which is of course equal to the interest rate, this casual usage has caused a lot of confusion. We therefore strongly suggest that "positive time preference" be defined to exclude the effect of a difference in marginal utility due to any lower level of consumption in the present and, therefore, to include only any preference for present over future consumption due to other causes. In our formulation, time preference is then defined as a positive value of the constant η, rather than in terms of the entire expression (3), which gives the slope of the indifference curve. The ratio of marginal utilities, $[v'(C_0)]/[v'(C_1)]$, depends not on time but on levels of consumption in the two periods; any preference for present over future utility, by contrast, can be conveniently represented by the subjective discount rate η.

II. Separable Utility

Before we can properly derive our positive results, we need to consider whether the view of intertemporal choice set out in equation (1)

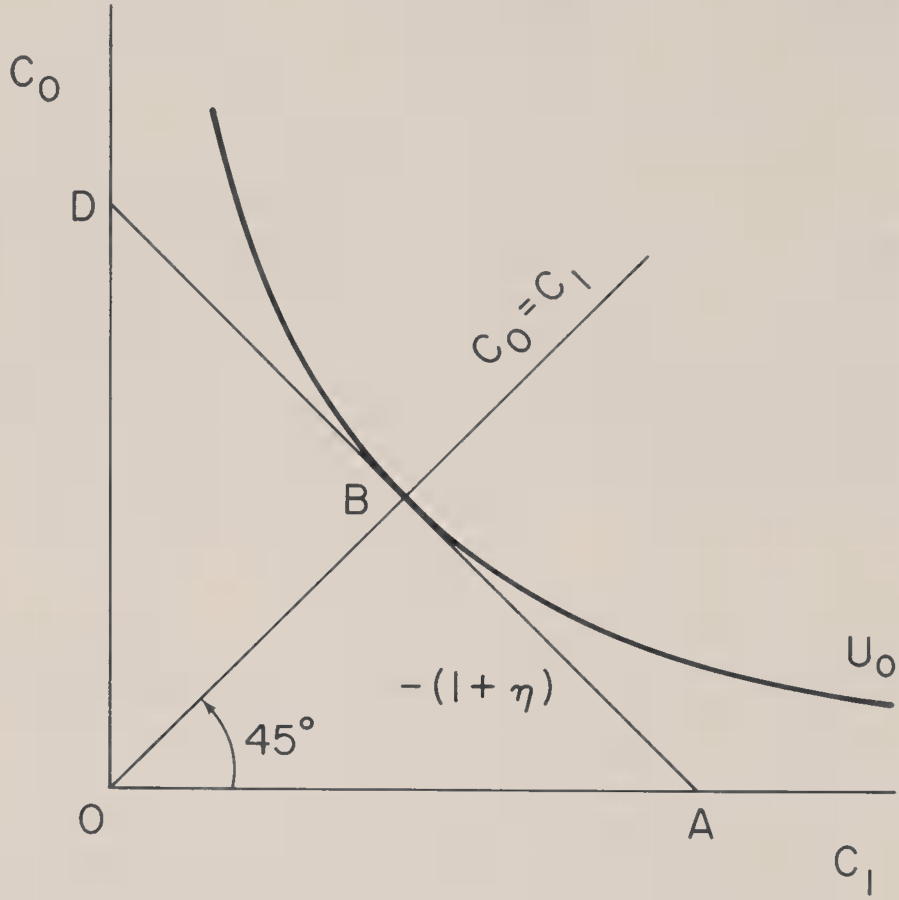


FIG. 1

is appropriate. In particular, we need to consider whether it is necessary or expedient to write utility as a function of consumption in every period:

$$U = U(C_0, C_1, C_2, \ldots, C_T). \tag{6}$$

This formulation, unlike that in equation (1), would formally encompass any complementarity between the levels of consumption in different periods. Indeed, one could go further and write utility as a function of the amount of each separate good consumed in each period—as would be necessary, for example, if the typical household consumed a little cream today to go with the coffee it will consume a week from Monday. In other words, the utility derived from a given level and pattern of consumption in a particular period could conceivably depend not only on the level but also on the composition of consumption in prior periods. There are examples of interdependence of consumption on different dates, such as a preference to have different dinner menus on successive days, and it could be necessary for some special purposes to take these into account. However, as

Samuelson (1937) has argued, functions that allow unlimited interrelationships become so general as to be almost vacuous. If a utility function rules nothing out, it will be consistent with almost any possible type of behavior and thus generate no potentially falsifiable hypotheses of interest. In addition, it has been shown elsewhere (Bailey and Olson 1977; Bailey et al. 1980) that the assumption of separability is more realistic than is usually supposed. Finally, it is hard to imagine how any plausible type of temporal complementarity could generate results that are qualitatively different from those in this article, with one obvious exception we can handle now.

An individual cannot, of course, derive utility from consumption in any period unless he has survived prior periods and also preserved the health needed to enjoy consumption. This could be accommodated formally by treating the essentials needed to preserve life and health as crucial investments in human capital. This would not really solve the problem, though, since the temporal interdependence would still have to be taken into account in investment decisions. Similarly, if a family is to have grandchildren or more distant descendants, then each intermediate generation must have at least enough resources to survive long enough to raise children. Thus we emphasize that our analysis must be modified before it can be applied in any situations where levels of consumption too low for healthful viability or reproduction are at issue. Abstractly, we deal with this starvation level of consumption ${}_0C$ by supposing that $v'(C) \rightarrow \infty$ as $C \rightarrow {}_0C^+$ and that $v(C)$ is undefined for $C < {}_0C$. Given this qualification, we can safely continue to use function (1).

III. Erroneous Inference of Myopic Behavior

In Section I we distinguished time preference from the effects of diminishing marginal utility of consumption in each period. With the aid of this distinction we can clear up one elementary yet surprisingly common error. Since Böhm-Bawerk's time, if not before, some economists have assumed that positive interest rates indicate that most people have positive time preference. Though we shall later provide evidence that time preference is in fact positive, this finding by no means follows from the existence of interest.

This is obvious once one considers the implications of the fact that capital goods can be useful in production. The existence of profitable investment opportunities implies that reduction of consumption by a dollar now will allow more than a dollar's increase in consumption in the future. A glance at equation (1) reminds us that even if time preference (η) were zero, this fact can generate positive interest rates, because diminishing marginal utility of consumption in each period

entails that consumers will accept a lower level of consumption in one period than another only if they are rewarded for doing so. Obviously, positive interest rates and the excess of marginal utility of present over that of future consumption that they induce[3] does not by itself justify the conclusion that there is any preference for present over future utility, much less persistent myopia or irrationality.

When one looks at both the demand and supply sides, it becomes clear that each of the following combinations of conditions is sufficient to generate the customary positive interest rates: (1) diminishing marginal utility plus profitable investment opportunities with nondecreasing income (and zero time preference); (2) diminishing marginal utility with increasing endowment income (and zero time preference); (3) positive time preference with constant income or with profitable investment opportunities (and nonincreasing marginal utility). Therefore, the fact that interest rates are positive does not by itself tell us which causes are operative. Only evidence of a different kind could demonstrate that there is positive time preference.

The rudimentary considerations just mentioned are, however, sufficient to tell us something about the relationship of η, whether it be positive, negative, or zero, and the prevailing interest rate. Consider initially a manna-from-heaven world with no producers. Such a world could have a consumption-loans market with a determinate rate of interest in long-run equilibrium. Since we have not specified the times at which manna falls from heaven, or yet offered any evidence that would either entail or rule out positive time preference, the interest rate in this consumption-loans market could be anything—even zero or negative. If the manna falls evenly over time, the interest rate in this consumption-loans market is bound to be equal to η.

But now suppose that advantageous production processes are discovered in which produced capital is an argument in the production function for consumption goods. Since this discovery will increase the demand for borrowings but not shift the supply function for loans, the interest rate must rise. The result is that in equilibrium $r > \eta$. Except for a few extreme cases, the real world will be one in which $r > \eta$, because in almost all real-world cases capital goods have a positive net marginal product. The marginal rate of substitution between present and future consumption in long-run equilibrium will of

[3] The difficulties involved in any effort to get one unambiguous measure of the capital stock and the extended controversies that have grown out of this should not, of course, obscure the fact that capital accumulation requires postponement of consumption. Böhm-Bawerk's writings, despite their well-known shortcomings, are sufficient to show this (see Dorfman 1959*a*, 1959*b*). Thus the presence of "reswitching," e.g., does not undermine our analysis here.

course always equal the rate of interest, but η is not in general equal to this marginal rate of substitution, even if η is positive.

IV. Is Time Preference Observable?

It is intuitively obvious from the discussion of figure 1 and from what has just been said that we can ascertain an individual's rate of time preference by observing the interest rate at which he chooses a level consumption stream. This means that we have a firm and simple basis for using a revealed preference approach to obtain solid scientific evidence about time preference and have no need to appeal to anything resembling introspection or interpersonal comparisons of utility.[4]

The conclusion that the household chooses a level consumption stream if the rate of interest equals the rate of pure time preference clearly holds for the multiperiod as well as for the two-period cases (see Samuelson 1937). This finding is itself enough to make time preference observable. We noted earlier that, unless incomes are declining, the existence of productive investment opportunities implies that the equilibrium rate of interest will be greater than the rate of time preference, and this is true however high the rate of time preference might be. Therefore we would predict that in most cases households would have positive levels of savings that generate a rising level of consumption over time. This theoretical prediction appears to be relatively realistic.

There are also other ways to measure time preference from revealed preferences. If the household maximizes utility (1) subject to the budget constraint

$$\sum_{t=0}^{T} \frac{Y_0^* - C_t}{(1 + \bar{r})^t} = 0, \tag{7}$$

varying each C_t to obtain this maximum, the result implies

$$\frac{v'(C_t)}{v'(C_0)} = \frac{(1 + \eta)^t}{(1 + \bar{r})^t} \tag{8}$$

[4] In seeking revealed preference we are, of course, only continuing a line of inquiry started by others. In 1937 Samuelson proposed a hypothetical method for obtaining the utility function by integration, i.e., by mapping observed choices into a functional representation; in effect this was applying the revealed-preference approach to our problem. Similarly, Debreu (1959) showed, with a rigorous proof, that a "preordering," or a consistent, transitive set of preferences, implies a utility function that is unique up to a linear transformation. Frisch's (1959) analysis also proceeded in a similar spirit. There are also others who have made significant contributions to this way of researching the problem, such as Strotz (1956) and Hirshleifer (1970), whose contributions will be discussed later.

with a constant interest rate $\bar{r}$ for all t. Evidently, as we noted above, $C_t = C_0$ when $r = \eta$, so that when we find the rate of interest r_0 at which the household chooses a level consumption stream (i.e., zero savings from Y_0^*), then $\eta = r_0$, and the value of η has been revealed.

Similarly, if $r > \eta$, as it usually is, $C_t > C_0$. If one has independent information on how rapidly $v'(C)$ declines with increasing C (see Bailey et al. 1980), one can estimate η simply by examining the chosen growth path of consumption and finding the η that reconciles it with equation (8).

V. Uncertainty

Though there are some notable exceptions, such as Ramsey (1928), and more recently Stigler and Becker (1977), most economists appear to take it for granted that time preference is positive. Sometimes this conclusion seems to rest not so much on the erroneous inferences we described in Section III as on the notion that the individual is uncertain about the date of his own death and will therefore discount prospective utility because of the possibility he will die earlier than expected. This argument is quite unsatisfactory for at least two reasons.

First, uncertainty about the date of one's death can increase as well as decrease an individual's provision for the future, because it can create an incentive to save more to provide undiminished consumption in the event of an unexpectedly long life. If income could costlessly be converted into an annuity, this would not be the case, but ideal annuities are not available, and in less developed economies they often cannot be purchased at all. It is an empirical question whether uncertainty about the length of one's life increases or decreases provision for the future.

A second reason why uncertainty about the date of one's death does not entail positive time preference is evident from the fact that there is normally positive saving in the society as a whole. Even if there were negative time preference over the life cycle, that is, a preference for utility in later as against earlier years of life, a society with no population growth or technological advance would have no social saving whatever if everyone consumed his entire output over the course of life: The savings in one set of years would just equal dissaving in another. In fact, the developed societies have saved enough to provide for a considerable long-term growth of tangible capital per workers as well as a vast increase in the amount of education and human capital. This substantial social saving had to be due in large part to a desire to bequeath increased consumption opportunities to descendants. The desire to bequeath assets to descendants provides

another reason why the argument that uncertainty about the date of one's death implies positive time preference is inadequate.

There is also uncertainty about whether the assets obtained through savings will generate the promised yield or even preserve their value. Such risks lower certainty-equivalent rates of return, and so are different and distinct from time preference. Risks may be reduced through diversification of portfolios, but transactions costs limit the extent to which risk can be eliminated in this way for the household of average means, and there are in addition great problems for such a household in diversifying against the risks to property in any single country (or single set of countries that might have a common destiny). Thus the risk that investments will prove unrewarding would appear to reduce saving, just as a positive rate of pure time preference would.

Uncertainty nevertheless also has two important countervailing effects that have been neglected, if noticed at all, in the literature. The first of these arises from risk aversion and from the absence of futures markets and insurance markets for most goods. The absence of futures markets makes it impossible for a typical household to hedge against a loss of demand for whatever specialized skill it has to offer, or against the loss in value of whatever other assets it possesses, or against a depression. Similarly, a typical household cannot insure against the birth of a child that will never be capable of supporting itself or even fully insure against a loss of the skill or health needed to generate income. These unhedgeable and uninsurable risks to the total income-earning capacity of a household, in combination with risk aversion, entail an increased incentive to save. In other words, even uncertainty about the security of assets can work to increase as well as to decrease savings, because in unfavorable contingencies assets contribute a large increment to prospective utility. This implication is corroborated even by everyday language, as when people speak of "saving for a rainy day." Second, the greater the amount saved, the smaller in general will be the ratio of the transactions costs of diversifying the porfolio or protecting it in other ways[5] to the value of the portfolio. In summary, uncertainty may reduce the expected apparent rate of return to saving, yet at the same time in the absence of universal futures and insurance markets it can increase the "need" or incentive for saving, that is, its prospective utility. Because, in addition, uncertainty may favor accumulation of enough wealth to

[5] In the case of households whose assets are acquired by illegal or unpopular means, or which become so large they generate notoriety, additional saving may if anything increase the possibility of prosecution, extortion, or adverse political decisions. Yet concealment and diversification can reduce these risks also, perhaps more easily as wealth grows larger.

make diversification or other forms of portfolio protection practical, it need not lower even the apparent rate of return.

The question of whether those uncertainties that imply the same behavior as positive time preference are quantitatively more significant than those that imply the same behavior as a negative time preference is an empirical question. Yet to the best of our knowledge it has not even been addressed. Nor are we aware of any persuasive empirical literature on whether other possible influences on saving behavior, such as irrationality or perceptual bias, combine with the two opposing effects of uncertainty to create an appearance of time preference that is on balance positive, or zero, or negative.

VI. Is Time Preference Positive?

The argument in prior sections helps us address this question now. We shall see in the next section that there are reasons for assuming that some households can have an indefinitely long or infinite time horizon, and that the conclusion we shall draw with this assumption holds a fortiori when there are shorter planning horizons. If $\eta = 0$, from equation (8) we have

$$\frac{v'(C_t)}{v'(C_0)} = \frac{1}{(1 + \bar{r})^t} \,. \tag{9}$$

If we then extend the time horizon to later and later t and apply (9) to the "last" period in the future, we obtain

$$\lim_{T\to\infty} \frac{v'(C_T)}{v'(C_0)} = 0. \tag{10}$$

This result implies that with the continuing positive interest rates that normally prevail and an η of zero, the household would lower C_0 to the desperation level, where $v'(C_0) = \infty$, if the C_T attainable under the budget constraint (7) is below the satiation level, at which $v'(C_T) = 0$. Inasmuch as we normally do not observe either people who cut consumption down to utterly abject levels in order to provide for the future or people who have because of past saving reached consumption levels at which their taste for goods is satiated, there appears to be a positive rate of time preference, that is, in our formulation an $\eta > 0$. (We discuss later a possible escape from this conclusion if there is *both* an infinite time horizon and what we call "drastically diminishing marginal utility.")

We have shown earlier that r is normally greater than η because capital is useful in production, and the present argument suggests that if r exceeded η by a very large amount that would also generate implausibly low levels of present consumption. This suggests that, at least in those societies where real interest rates are very high and

expected to remain so, η must exceed zero by more than a miniscule amount.[6]

Before turning to finite planning horizons, we should examine a familiar policy conclusion which assumes an infinite time horizon. This is the prescription of those who claim or imply that, at least when ecological systems or depletable resources are at issue, the only right and moral attitude is, in effect, a zero rate of time preference and an infinite time horizon—the interests even of distant generations should have the same weight as one's own interest in current-year consumption. One need not have a what-has-posterity-ever-done-for-me attitude or a shortsighted view to oppose this supposedly moral command. It implies, as a glance at (10) will remind us, that at every positive interest rate the household should reduce its consumption to the level at which the marginal utility of current consumption is infinite, in order to give all to the future. Of course, that applies to every future year, and to every future generation, until either income rises so high that the marginal utility of consumption (when consumption equals income) has fallen to zero, or alternatively there is so much capital around that the interest rate is no longer positive. We doubt that the proponents of a zero time preference and an infinite time horizon have understood the implications of their argument and conclude that their argument surely cannot be a proper basis for national policy.[7]

The existence of positive time preference also has immediate implications for the responsiveness of savings to changes in interest rates. We earlier defined η as zero for a household if, when it had equal endowments in each period and faced an interest rate of zero, it chose a stable level of consumption. A positive η in these conditions, or any η greater than the interest rate, implies a decreasing level of consumption through borrowing or dissaving, whereas an $r > \eta$ implies net saving. If $\eta = r$ at any level, there is a stable level of consumption. Starting from such a situation and looking at equation (8), we ask what would happen if the interest rate fell to zero. The analogy with the results in equations (9) and (10) immediately reminds us that this would give us

$$\frac{v'(C_t)}{v'(C_0)} = \frac{(1+\eta)^t}{1}. \tag{11}$$

As T gets indefinitely long, this would give

$$\lim_{T\to\infty} \frac{v'(C_T)}{v'(C_0)} = \infty. \tag{12}$$

[6] We are thankful to Peter Murrell for calling this point to our attention.
[7] For a related finding, see the Koopmans (1960) article discussed below.

We see that either C_0 increases until $v'(C_0) = 0$, or $v'(C_T)$ decreases until $v'(C_T) \to \infty$. That is, the ratio C_0/C_t will become huge: The family will dissave on a large scale, unless it somehow suffers a permanent plight so severe that $v'(C_t)$ goes rapidly boundless with a small reduction in C_t, or so near satiation of the capacity to enjoy goods that $v'(C_0) \to 0$ quickly. This shows that it is logically contradictory to hold that the representative household is myopic or for any other reason has positive time preference, and also to hold that savings do not respond to changes in the interest rate and would normally remain positive at a zero interest rate.[8] (In fact, saving would be sensitive to the rate of interest even if time preference were negative and, since the marginal utility of consumption is normally diminishing, also if it were zero in any environment in which endowments vary over time.) We think this implication may surprise many readers. It will be clear from the next two sections that this conclusion, like the others, also holds true with finite time horizons.

VII. Truncated Time Horizons

Most significantly, the case that there is positive time preference—that is, something beyond Böhm-Bawerk's "first cause"—gets stronger as the planning horizon gets shorter. The reason for this is utterly obvious once stated: A finite planning horizon of T years implies that whatever happens after T years is not only discounted but given a value of zero. Were there not usually markets in which one can trade with those whose time horizons extend farther into the future, an individual would place no value at all on an asset, such as a bond, that could be redeemed or sold only after the end of his planning horizon. Truncating a decision maker's time horizon can therefore only strengthen the evidence for a positive time preference, since it merely involves attributing a zero value, rather than a discounted but positive value, to any consumption after a certain date.

The fact that truncated time horizons lead so directly and simply to positive time preference makes one ask how common and how long truncated time horizons might be. Are observed savings rates, which as we have shown are far lower than a zero time preference and an infinite time horizon would imply, as low as they are simply because the behavior we actually observe is driven by time horizons on average about a lifetime or so in length?

The notion that time horizons might average about a lifetime in

[8] We have in mind a compensated saving schedule, along which household real income is constant, so that the effect of a change in the interest rate is a pure substitution effect.

length has some appeal at first glance. It certainly must be an overestimate for older people without heirs or other survivors about whom they are concerned. Probably it is about right for some others too, who for one reason or another may feel that the cares of one lifetime are enough. The number with concern for heirs is probably larger, but it appears that mainly this concern extends only a generation or two into the future. Even in dynastic families, such as that of the Capetian kings of France, we may wonder whether they thought ahead much beyond producing a male heir and preserving his birthright. Taking the different types of cases together, it might seem that truncated time horizons might average out to perhaps a lifetime, and that this is the main explanation of the positive η that we inferred from the small percentage of incomes that are saved.

This interpretation is nonetheless unsatisfactory, as we realize from our earlier reference to the fact that lifetime horizons imply no saving by societies, unless the age distribution for a time contains disproportionate numbers of people in the "savings" stages of life. There is substantial social saving, especially when we include education and human capital as we must, that is not explainable by growth, so the typical time horizon must be more than a generation in length.[9]

This consideration in turn means that there is no special need to know the rate of time preference within planning horizons of a lifetime or less in length. Though individuals with time horizons as short as this presumably exist, the evidence adduced above suggests that they are exceptional. It is conceivable that these exceptional individuals could have a zero or negative time preference over the course of their planning horizons. If so, there is a remarkable discontinuity, for we know from the definition of a finite planning horizon that nothing that occurs after it is even worth taking into account. The dominant fact is that the time preference for consumption within the planning horizon is infinite compared to anything that happens after it. Whether one is interpreting aggregate time series on saving and accumulation or analyzing social policy about them, this infinite time preference is the dominant fact, besides which the presence or absence of time preference within the life span pales to insignificance.

It is indeed fortunate that time preference within planning horizons of a lifetime or less in length is not the key issue, for (contrary to what we once believed) it is almost impossible to find out what time preference over the life cycle might be. Family size and optimal

[9] Growth due to technological advance in what would otherwise be a steady state also implies some net social saving, because young savers have higher lifetime incomes than have older people, so that their total saving exceeds the dissaving of retired people. However, White (1978) and Kotlikoff and Summers (in press) find that total saving far exceeds the amount that can be attributed to life-cycle saving in a growing economy.

periods for major investments in human capital change over the course of life, and though it is easy to incorporate these considerations into our theoretical framework, the changes in spending and savings levels they bring about are difficult to disentangle from any short-run time preference. Even the individual's capacity to enjoy consumption may vary with age, as is suggested by the cliché that one should not postpone high levels of consumption until one is "too old to enjoy it." In most households the wage rate also changes substantially with age, and this again affects the level of savings in each period. There are even problems arising because of changes in the value of time due to these changes in wage rates. The final outputs desired by consumers are characteristically produced through household production functions in which time and market goods are inputs. As Becker and Ghez (1975) have shown, changes in wages over time can affect savings levels in at least two significant ways. First, the higher the wage rate the greater the incentive to substitute purchased goods for time in the domestic production of final outputs, as may occur when a family increases its reliance on caterers, convenience foods, and restaurants when wage rates of family members are unusually high. Second, the anticipation of a different opportunity cost of time in the future can change the time profile of final consumption and savings, so that there is more consumption in periods when a lower cost of time makes the price of final output cheaper, as could occur when a family postpones the spending involved in distant tourism till retirement. Given the unknown and sometimes contrary effects of these and other influences, it is not immediately clear what levels of savings would support or refute the hypothesis that time preference is positive within time periods of a lifetime or less.

VIII. Bequests

Let us now return to the more representative case of a household with heirs for which it has some concern. What a couple during the course of life gives to its children and what it bequeaths to them has a capital value equal to the discounted future value of the consumption (and later bequests) it is expected to provide. In the minds of the givers this bequest must have a prospective value or utility at least equal to the utility forgone by not consuming it themselves. The extent of any willingness of the household to reduce the heads' current consumption to provide for their children's endowments is the central parameter revealing the presence or absence of the "myopia" discussed in debates about the determinants of saving.

A household with the faith that it will have a continuing line of descendants with the same regard for their children as it has for its own will act as if it had an indefinitely large or infinite time horizon, even if it gives such a long time horizon no explicit attention. Suppose that the household spends that amount on education, gifts, and bequests that strikes its preferred balance between its own desire to consume and its provision for the children's future, and leaves it to the children to allocate their resources in turn. When the problem is understood in this way it becomes clear that there is an extensive theoretical literature that is directly applicable, notably that of dynamic programming. One of the basic findings in this literature is that the optimal program for several periods is the same as that reached by optimizing each period in turn, including setting optimal end-of-period conditions each time for the beginning of the next period (see, e.g., Bellman and Kalaba 1965, chap. 2); the result is the same whichever way one looks at the problem. That this notion applies precisely to intergenerational resource allocation is shown by Barro (1974). It can accordingly be appropriate to view the choice of the size of the bequest as if the household were maximizing a utility function of the form (1) with an indefinitely large, or infinite, time horizon. The T represents not how far into the future people plan consumption, but the length of a dynamic program implied by a continuing series of intergenerational transfers.

Let us now suppose that the conditions that make it appropriate to assume an infinite time horizon do not hold, so that finite time horizons such as we used to develop equations (1)–(8) are appropriate. If the present heads of household are utterly unconcerned about any descendants beyond, say, their grandchildren, this lack of concern will manifest itself in the form of a smaller bequest than otherwise. Uncertainty about whether a subsequent generation will have offspring, or fail to make a bequest to its children, will also make for a smaller bequest. The effect of a lack of concern for more distant generations, or of uncertainty about whether the bequest will in fact serve the givers' purposes, is approximately the same as if the household applies a high rate of time preference η in maximizing utility (1) with a large or infinite T. We can see this by taking equation (1) and making $v(C_t)$ a declining function of time, for example, $v(C,t)$ with $v_t < 0$. This is in effect equivalent to keeping (1) as it is and raising the value of η. This makes it clear in another way that less-than-infinite time horizons entail positive time preference. Thus the decision to let T in equation (1) be finite at some points, and to let it approach infinity in Section VI, was governed largely by analytical convenience and had no substantive effect on the conclusions.

IX. Drastically Diminishing Marginal Utility

With infinite time horizons and continued positive interest rates the returns to savings ultimately become so colossal that, if there were not positive time preference, current consumption would be drawn down to survival levels. There are, however, two phenomena in addition to positive time preference that conceivably could explain the high levels of consumption that actually occur. One is that such large increases in endowment income are expected that saving for still more future consumption is not worthwhile however much interest would accrue.[10] The other possibility is that the marginal utility from consumption in any period diminishes so drastically as consumption rises that even essentially unlimited interest income would not yield enough future utility to justify much saving. There is no reason for most households to assume increases in endowments so great that they could justify the low observed savings rates. It is, on the other hand, just possible that drastically diminishing marginal utility in combination with plausible expectations about increasing endowment incomes, perhaps due to technological advance, could explain the observed levels of saving.

Suppose, for example, that the household expects an exponentially rising real income due to technological advance, even if its saving and aggregate savings are zero. Equation (8) implies a chosen rate of growth of consumption that is independent of the horizon T; this growth will be exponential if marginal utility is a constant-elasticity function, that is, if utility has the form

$$v(C) = AC^{\alpha} + B \tag{13}$$

for $\alpha < 0$, $A < 0$, and $B > 0$, or, for the unit elastic case, the form

$$v(C) = A \ln C + B \tag{14}$$

with $A > 0$ (see White 1978). For Harrod-neutral technological advance, there is some fraction of income saved that is consistent with having consumption grow exponentially at the same rate ρ as income does. If so we can have $v'(C_T) \to 0$ as $T \to \infty$ because of this growth.

For example, suppose that utility is given by the Bernoulli function (14). Then the optimum consumption pattern implied by (8) is

$$C_t = \frac{(1 + \bar{r})^t}{(1 + \eta)^t} C_0 \tag{15}$$

for every t. If it should happen that $1 + \rho = (1 + \bar{r})/(1 + \eta)$, or, approximately,

[10] We are grateful to Christopher Clague for calling our attention to this possibility.

$$\rho = \bar{r} - \eta, \tag{16}$$

consumption will grow at the same rate as income and will remain a constant fraction of income, as is actually observed. This pattern can be sustained indefinitely and can plausibly be reconciled with a constant rate of interest (see Solow 1970, chap. 5).

If the rate of time preference were zero, equation (15) would imply that consumption rises indefinitely at the rate $\bar{r}$. When $\bar{r} > \rho$, as has been true of the United States, this high rate of growth of consumption can be maintained for large T only by pushing C_0 down toward zero, pushing the fraction of income saved toward unity.

This set of results and considerations makes a compelling case for the existence of a positive rate of time preference, unless one supposes that the utility function $v(C)$ is inelastic and that unlimited exponential growth of income is expected even with zero saving. In the past, output per unit of input has grown at about 1 percent a year, and the U.S. birthrate has converged on zero population growth. If the real rate of interest of about 5 percent (which has prevailed during most of the past several decades) is what households expected to continue into the indefinite future, these figures can be shown to reconcile with zero time preference only if the elasticity of the utility function (13) is minus 5 corresponding to $\alpha = -4$. This function has what we shall label "drastically diminishing marginal utility." We do not argue that utility functions of this character are the norm but know of no data that rule them out and think it is not without interest that (when exogenous increases in endowments are expected and time horizons are also infinite) the facts of saving force us to choose between positive time preference and drastically diminishing marginal utility of income.

X. The Relationship to Prior Findings

Though many of the prior findings on time preference contradict one another, and there has been no generally accepted analysis of the matter, many of these writings nonetheless have considerable value. With the aid of the findings here we are in a better position to understand or appreciate some prior contributions.

One of the best-known contributions is that of Strotz (1956). The extreme simplicity of Strotz's insight and its relationship to our argument can be made immediately apparent. Suppose that an individual has a positive time preference, in that he values consumption in the future less highly than consumption now, but it is only the distance into the future, rather than the specific date, that determines the weight given to the utility. Suppose for the sake of a simple example that the marginal utility of consumption in any one period is the

same at all levels of consumption, however high or low, so that we can ignore the diminishing marginal utility of consumption and focus exclusively on Strotz's insight. Let us suppose that our individual must obtain all consumption out of an initial endowment of 25 and that his time preferences are such that he chooses initially to consume 20 units, or four-fifths of his stock, in the present period, three units in the second period, and two units in the third. When the first period is over, however, the individual will abandon his initial plan and consume four-fifths of his remaining endowment of five, or four, since by stipulation he weighs the utility in the present so highly he consumes four-fifths of his stock, and period 2 is now the present. This is of course inconsistent with the initial plan for consumption in the second and third periods; the initial plan has not been followed, even though all of the individual's expectations about his situations and tastes have proved correct. The point is that no individual with the pattern of time preference described will voluntarily adhere to the plans he makes in advance; it is impossible for him to maintain his pattern of time preference and also adhere to the intertemporal plans he makes at the beginning. It becomes obvious on reflection that the only patterns of positive time preference that individuals can have and still make consistent or rational intertemporal choices are those which involve a constant exponential rate of discounting: a constant η such as we have all along assumed. If an individual plans to save the same percentage of his period 2 stock as of his period 1 stock, he will in fact carry through with that plan if his tastes do not change. Of course, no positive time preference at all ($\eta = 0$) is consistent with rational behavior; it would have led our hypothetical individual to consume a third of his stock in each of the periods and to have adhered to that plan. Thus we see that it is not consistent to assume both rational behavior and positive time preference without also assuming that η is an exponent of constant value.

Unfortunately, Strotz goes beyond the insight just described to define "prudent" or "thrifty" behavior as any behavior which meets the standard of intertemporal consistency: An individual is arbitrarily defined to be "thrifty" if he discounts the future at a constant exponential rate, however high that rate may be. In keeping with his preoccupation with perceptual error, inconsistent choices, and the psychology of intertemporal choices, Strotz concludes that "consumer sovereignty has no meaning in the context of the dynamic decision-making problems" (1956, p. 179).

Though Strotz's disturbing conclusion is consistent with the rest of his argument, we can now readily see that it is not in fact justified. If a positive time preference, which he calls "myopia," were solely the result of irrationality and perceptual errors, then consumer sovereignty would indeed have little or no meaning for intertemporal

choice. But as we have seen, the absence of a positive time preference implies extreme behavior, or else drastically diminishing marginal utility. Short planning horizons can occur and need not imply any irrationality, and the very act of ignoring any return beyond some limited planning horizon as we know entails a positive time preference. Thus, though the intransitive and irrationally impatient behavior Strotz illuminated can of course also occur, his nihilistic general conclusion about the inapplicability of consumer sovereignty to intertemporal choices is incorrect.

Our analysis is in closer accord with writings of Koopmans, Diamond, and Williamson (1964) and Hirshleifer (1970). These writers also conclude that there is positive time preference, which they define somewhat differently. They do not, however, offer any "empirical test" of the sort we used, nor go into shorter time horizons and drastically diminishing marginal utility. Koopmans et al. (1964) prove that various sets of formal postulates entail the existence of positive time preference. The formal postulates are naturally chosen in part because of their tractability, and it is not always immediately clear whether these postulates are very restrictive or succeed in capturing the essence of the matter. It is not immediately clear, for example, whether Diamond's (1965) assumption that each period's utility $v(\cdot)$ is bounded is limiting, or even material to the proof, or whether it matters that these authors exclude the possibility that $v'(C_0) = \infty$ or that $v'(C_t) = 0$, which we allow. The fact that more work needs to be done before the generality and applicability of the Koopmans (1960) and Diamond (1965) proofs can be definitely established is perhaps best demonstrated by quoting Koopmans's briefest summary of the reason for positive time preference in his article: ". . . if there is in all circumstances a preference for postponing satisfaction—or even neutrality toward timing—then there is not enough room in the set of real numbers to accommodate and label numerically all the different satisfaction levels that may occur in relation to consumption levels for an infinite future" (1960, p. 288). Hirshleifer (1970, p. 96), using a line of reasoning much closer to ours, concludes in an important if cryptic passage that zero time preference would imply infinite deferral of all consumption. That is closer to the mark but overlooks some of the possibilities. In general, if the argument in our paper is correct, it suggests that the contributions we have just cited deserve more attention and perhaps further extension as well.

More recently, Stigler and Becker (1977) have put forth conclusions that directly contradict those of Strotz (1956), Koopmans et al. (1964), and Hirshleifer (1970). Stigler and Becker argue that many types of behavior that economists have traditionally characterized as special patterns of taste are in fact explicable in terms of different prices and incomes. One of the ad hoc assumptions about tastes to which Stigler

and Becker especially object is the assumption, going back as we know to Böhm-Bawerk, that people "'systematically undervalue . . . future wants.' The taste for consumption in say 1984 is alleged to continue to shift upward as 1984 gets closer to the present. In spite of the importance frequently attached to time preference, we do not know of any significant behavior that has been illuminated by this assumption. Indeed, given additional space, we would argue that the assumption of time preference impedes the explanation of life cycle variations in the allocation of resources, the secular growth in real incomes, and other phenomena" (Stigler and Becker 1977, p. 89).

It might seem that the quoted passage by Stigler and Becker contradicts our own findings no less than those of the authors we have cited, since we also claim to have demonstrated the existence of positive time preference. This is not in fact the case; one could consistently accept Stigler and Becker's conclusion and our own as well. We have defined positive time preference in a slightly broader and more readily testable fashion than Stigler and Becker have, and we neither assumed nor ruled out irrationality or perceptual error. Accordingly, it is, as we argued, an empirical question whether there is time preference, and the revealed preference approach made it possible to assess the evidence. It would of course be possible to discuss the question of why bequests are as small as they are, or why some people have only finite planning horizons, with different terminology, though excluding the phrase "time preference" and synonyms for it would in our judgment involve circumlocution.[11]

XI. Conclusions

Our principal findings are the following:

1. A finite time horizon implies that only consumption inside the horizon has value and signifies an absolute unwillingness to transfer goods to children or other descendants who could benefit from them after a specified future date. Thus a finite planning horizon is both theoretically and practically equivalent to a positive rate of pure time preference. Planning horizons that do not reach beyond the lifetimes of those who make the plans can exist, and (though this implies a remarkable discontinuity) it is logically possible that within such lifetime horizons time preference could be zero or even negative. But this possibility can have little significance in comparison with the high rate of time preference implied by the short horizon, and in any event it does not seem feasible to disentangle time preference during

[11] Stigler and Becker's footnote assertion that "a consistent application of the assumption of stable preferences implies . . . the absence of time preference" (1977, p. 78, n. 4) is either mistaken or dependent upon definitions of "stable preferences" or "time preference" that appear to be untypical.

the course of a lifetime from other changes over the life cycle. Time horizons which imply no bequests for descendants also cannot be the norm, since they would imply lower savings than are observed and no social saving whatever in a society in a stationary state. We have also shown that finite time horizons and positive time preference are functionally equivalent, so the most interesting question is whether there is positive time preference within an infinite time horizon. If there is, there is no need to go into the length of planning horizons, and the analytically convenient assumption of an infinite time horizon can be used without significant loss of generality.

2. An infinite time horizon and a zero rate of time preference suggest that (if the foreseeable rate of economic growth is less than the expected long-term rate of interest) current consumption would be no higher than the subsistence level. Over an indefinitely long time span continued positive interest rates imply such colossal gains in future consumption from reductions in current consumption that all income beyond that needed for survival would be saved.

3. There is, however, the alternative possibility that even vastly higher consumption is expected to bring so little increase in utility that households are unwilling to forgo consumption now, even though this can generate essentially limitless increases in consumption later. The possibility that there will be absolute satiation at these future consumption levels can be rejected on the classical grounds that wants are limitless, but there remains the intriguing possibility that there is what we here labeled "drastically diminishing marginal utility": Even boundless increases in future consumption would then bring such modest increases in utility that the low levels of saving we observe could be rational even with a zero rate of time preference. The observed levels of saving imply positive time preference, or drastically diminishing marginal utility, or both of these interesting phenomena. In an age that has seen secularly rising incomes, it might seem that drastically diminishing marginal utility is a type of positive time preference, since an unwillingness to bequeath much to descendants, on the grounds that they will be better off than we are and will not get much satisfaction from still higher consumption, provides a good rationale for preferring current consumption and leads to the same observed savings levels. This is not in fact correct. If there were a sufficient technological regression or other losses in endowments, yet a positive interest rate, zero time preference would imply that current consumption would be reduced to subsistence levels, but drastically diminishing marginal utility would not. Similarly, if for shifts of goods within their own lifetimes people behave as if they had Bernoulli utility functions, intertemporal consistency would require the use of positive time preference to reconcile this behavior with their gifts and bequests. Also, whether drastically diminishing utility is

present or not, finite planning horizons entail positive time preference. Drastically diminishing marginal utility also may be needed to explain other phenomena, such as a willingness to pay a high price to insure against even modest losses or to reduce fluctuations in consumption levels, and savings levels in particular cases can be traced to drastically diminishing marginal utility according as its other implications are present or absent.

4. Uncertainty about the future has double-edged effects. Of course, if the typical household thinks that there is a greater likelihood that part of its bequest will merely be destroyed in a future catastrophe than that the full bequest adds to the probability that descendants will survive well in that catastrophe, the expected rate of return to saving is thereby lowered in their minds. That expectation, instead of time preference, could in principle account for high consumption and low saving. However, each type of uncertainty we can think of offers possible reasons for saving more as well as reasons for saving less. As in the other cases, it is hard to see a simple way to distinguish it empirically from pure time preference. Thus for purposes of this inquiry we have made the simplest assumption: that uncertainty about whether any marginal savings will be lost is on average about offset by the incentives for additional saving that uncertainty about the future creates. To make a better specification possible in any further studies, we have made clear above that the prevalence of futures and insurance markets, the degree of risk aversion, the degree of confidence in predictions about the future, and the costs of diversification or protection of portfolios each has distinct effects on savings rates, and these might eventually be estimated.

In summary, the case for positive time preference is absolutely compelling, except in the interesting case where there is drastically diminishing marginal utility. Utility functions of this sort have interesting implications of their own which need to be tested separately. Time preference is accordingly by no means a subject that deserves only casual references to alleged perceptual errors or erroneous inferences from the existence of positive interest rates. It is a phenomenon of great practical and theoretical importance that is fortunately quite amenable to analysis.

References

Bailey, Martin J., and Olson, Mancur. "Pure Time Preference, Revealed Marginal Utility, and Friedman-Savage Gambles." Working Paper no. 77-2, Univ. Maryland, 1977.

Bailey, Martin J.; Olson, Mancur; and Wonnacott, Paul. "The Marginal Utility of Income Does Not Increase: Borrowing, Lending, and Friedman-Savage Gambles." *A.E.R.* 70 (June 1980): 372–79.

Barro, Robert J. "Are Government Bonds Net Wealth?" *J.P.E.* 82, no. 6 (November/December 1974): 1095–1117.

Becker, Gary S., and Ghez, Gilbert R. *The Allocation of Time and Goods over the Life Cycle.* New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1975.

Bellman, Richard E., and Kalaba, Robert. *Dynamic Programming and Modern Control Theory.* New York: Academic Press, 1965.

Böhm-Bawerk, Eugen von. *Capital and Interest.* Vol. 2, *Positive Theory of Capital.* South Holland, Ill.: Libertarian Press, 1959.

Debreu, Gerard. "Topological Methods in Cardinal Utility Theory." In *Mathematical Methods in the Social Sciences*, edited by Kenneth J. Arrow, Samuel Karlin, and Patrick Suppes. Stanford, Calif.: Stanford Univ. Press, 1959.

Diamond, Peter A. "The Evaluation of Infinite Utility Streams." *Econometrica* 33 (January 1965): 170–77.

Dorfman, Robert. "A Graphical Exposition of Böhm-Bawerk's Interest Theory." *Rev. Econ. Studies* 26 (February 1959): 153–58. (*a*)

———. "Waiting and the Period of Production." *Q.J.E.* 73 (August 1959): 351–72. (*b*)

Feldstein, Martin S. "The Social Time Preference Discount Rate in Cost Benefit Analysis." *Econ. J.* 74 (June 1964): 360–79.

Frisch, Ragnar. "A Complete Scheme for Computing All Direct and Cross-Demand Elasticities in a Model with Many Sectors." *Econometrica* 27 (April 1959): 177–96.

Hirshleifer, Jack. *Investment, Interest, and Capital.* Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Jevons, W. Stanley. *The Theory of Political Economy.* 5th ed. New York: Kelley, 1965.

Koopmans, Tjalling C. "Stationary Ordinal Utility and Impatience." *Econometrica* 28 (April 1960): 287–309.

Koopmans, Tjalling C.; Diamond, Peter A.; and Williamson, Richard E. "Stationary Utility and Time Perspective." *Econometrica* 32 (January–April 1964): 82–100.

Kotlikoff, Laurence J., and Summers, Lawrence H. "The Role of Intergenerational Transfers in Aggregate Capital Accumulation." *J.P.E.* (in press).

Marglin, Stephen A. "The Social Rate of Discount and the Optimal Rate of Investment." *Q.J.E.* 77 (February 1963): 95–111.

Pigou, Arthur C. *The Economics of Welfare.* 4th ed. London: Macmillan, 1960.

Ramsey, Frank P. "A Mathematical Theory of Saving." *Econ. J.* 38 (December 1928): 543–59.

Samuelson, Paul A. "A Note on Measurement of Utility." *Rev. Econ. Studies* 4 (February 1937): 155–61.

Solow, Robert M. *Growth Theory: An Exposition.* New York: Oxford Univ. Press, 1970.

Stigler, George J., and Becker, Gary S. "De Gustibus Non Est Disputandum." *A.E.R.* 67 (March 1977): 76–90.

Strotz, Robert H. "Myopia and Inconsistency in Dynamic Utility Maximization." *Rev. Econ. Studies* 23, no. 3 (1956): 165–80.

White, Betsy Buttrill. "Empirical Tests of the Life Cycle Hypothesis." *A.E.R.* 68 (September 1978): 547–60.

Temporary Income Taxes and Consumer Spending

Alan S. Blinder

Princeton University and National Bureau of Economic Research

Both economic theory and casual empirical observation of the U.S. economy suggest that spending propensities from temporary tax changes are smaller than those from permanent ones, but neither provides much guidance about the magnitude of this difference. This paper offers new empirical estimates of this difference and finds it to be quite substantial. The analysis is based on an amendment of the standard distributed lag version of the permanent income hypothesis that distinguishes temporary taxes from other income on the grounds that the former are "more transitory." This amendment, which is broadly consistent with rational expectations, leads to a nonlinear consumption function. Though the standard error is unavoidably large, the point estimate suggests that a temporary tax change is treated as a 50-50 blend of a normal income tax change and a pure windfall. Over a 1-year planning horizon, a temporary tax change is estimated to have only a little more than half the impact of a permanent tax change of equal magnitude, and a rebate is estimated to have only about 38 percent of the impact.

In 1968, faced with a classic case of demand inflation, Congress enacted a temporary increase in personal income tax payments to

This paper has benefited from seminar presentations at Princeton, Hebrew University of Jerusalem, Tel-Aviv University, and the University of Pennsylvania, and from helpful suggestions from Walter Dolde, Marjorie Flavin, Roger Gordon, Levis Kochin, Robert Solow, James Trussell, and a referee of this *Journal*. Most of the laborious computations were done expertly by Suzanne Heller; additional research assistance was provided by David Card, Robin Lindsey, and William Newton. Stephen Goldfeld and Richard Quandt generously guided this novice through the intricacies of their nonlinear optimization routines. Financial support from the National Science Foundation and from the Institute for Advanced Studies in Jerusalem, where part of this work was done, is gratefully acknowledged. This research is part of the National Bureau of Economic Research's program in economic fluctuations, but any opinions expressed herein are those of the author, not of the NBER.

[*Journal of Political Economy*, 1981, vol. 89, no. 1]

curb aggregate demand. In 1975, near the trough of our worst postwar recession, Congress enacted a tax rebate and other temporary decreases in taxes and increases in transfer payments designed to stimulate aggregate demand. Questions have been raised about the effectiveness of both measures.

The questions have both theoretical and empirical roots. On theoretical grounds, the permanent income–life cycle hypothesis seems to argue that temporary income tax changes should have little effect on consumer spending in principle. On empirical grounds, data on consumer behavior seem to suggest that the impacts of the two temporary taxes on spending were also small in practice. In 1968, the savings rate fell from 7.5 percent in the quarter immediately preceding the tax surcharge (1968:2) to only 5.6 percent in the first quarter of the surtax, suggesting that consumers kept spending despite the tax. In 1975, the savings rate ballooned from 6.4 percent just prior to the rebate to a stunning 9.7 percent in the quarter of the rebate (1975:2), suggesting that little of the rebate was spent.

The purpose of this paper is to study these two temporary tax changes in some detail. Precisely what prediction does economic theory make about the relative effectiveness of temporary versus permanent tax changes? And what conclusions can be reached from U.S. time-series data? The fact that the Carter administration asked for (but did not get) a repeat performance of the rebate in 1977 suggests that there is more than academic interest in the answers to these questions.

Section I outlines the theoretical issues, beginning with an idealized life-cycle model and proceeding to introduce some important "real world" considerations. Since the discussion shows quite clearly that the issue is an empirical one, Section II reviews previous empirical work on the subject very briefly. Section III explains the underlying basis of the empirical model of this paper, relating it to recent literature on rational expectations and the permanent income hypothesis (PIH), and then Sections IV and V show how this basic conceptual framework was converted into an operational empirical model. The estimates are presented and analyzed in Section VI, and Section VII summarizes the main conclusions.

I. The Implications of Theory: Pure and Impure

1. *The Pure Permanent Income–Life Cycle Theory*

As Eisner (1969) pointed out some time ago, the PIH casts doubt on the effectiveness of income tax changes that are labeled as temporary because such measures have only minor effects on permanent income.

To develop a theoretical benchmark for the marginal propensity to consume (MPC) that the PIH suggests might apply to a temporary tax, consider a rarefied world in which consumers with exogenous earnings streams select consumption paths to maximize lifetime utility. If capital markets are perfect, only the discounted present values of the earnings streams matter, so suppose all households earn a constant income y per year. Assume further that households differ only in age, a; that the real rate of interest is zero; and that the subjective rate of time discounting is also zero.[1] The question is, If income taxes are raised by z per capita for the period from $t = 0$ to $t = t_1$, how much less will consumers spend over this interval?

In answering this question, there are three population groups to keep track of. People who are "alive" (in the economic sense) at $t = 0$ and who live past the expiration of the tax suffer an income loss of $t_1 z$ over the period. If T denotes the length of life, then each such person of age a consumes a fraction $t_1/(T - a)$ of this loss during the years in which the temporary tax is in effect. Thus the change in consumption per capita is $\Delta C_1 = z t_1^2/(T - a)$.

Old people who are alive at $t = 0$, but die before $t = t_1$, lose only $(T - a)z$ in income. However, since they have MPCs of unity during the surtax period as a whole, their change in consumption per capita is $\Delta C_2 = (T - a)z$.

Finally, we must worry about people who are born between $t = 0$ and $t = t_1$. If a, a negative number between 0 and $-t_1$, denotes the age of such a person, and he lives for $t_1 + a < t_1$ years during the period $0 \leq t \leq t_1$, his income loss is $(t_1 + a)z$. Since he spends a fraction $(t_1 + a)/T$ of this income during the period $0 \leq t \leq t_1$, the change in his consumption is $\Delta C_3 = [z(t_1 + a)^2]/T$.

To derive the aggregate change in per capita consumption, weight these groups by the age distribution, considering the ages $-t_1 \leq a \leq T$. In the simplest case of a uniform age distribution, $f(a) = 1/T$, the change is

$$\Delta C = \int_0^{T-t_1} \frac{\Delta C_1}{T} da + \int_{T-t_1}^{T} \frac{\Delta C_2}{T} da + \int_{-t_1}^{0} \frac{\Delta C_3}{T} da.$$

Working out the integrals and dividing by the total income that is taxed away during the period ($t_1 z$ per capita), we obtain

$$\text{MPC} = \frac{t_1}{T}(\log T - \log t_1) + \left[1 - \frac{T}{2t_1} + \frac{(T - t_1)^2}{2Tt_1}\right] + \frac{t_1^2}{3T^2},$$

where the three terms show the contributions of the three different population groups.

[1] This is essentially the model introduced by Modigliani and Brumberg (1954).

To take a concrete example, suppose the typical lifetime of a household head as a household head is $T = 50$ years. Then, according to this formula, the MPC for a 1-year tax ($t_1 = 1$) is .09, while that for a 2-year tax ($t_1 = 2$) is .15.[2] The MPC = .09 for a 1-year temporary tax is only a rough benchmark representing the pure PIH, and there are a number of reasons why the theory probably systematically understates the responses of consumers to temporary taxes (see below), so we should not take this number too seriously. Still, there are two lessons worth drawing from this simple exercise—lessons that have often been forgotten in the temporary-tax debate.

a) Income gains and losses from temporary taxes will eventually be spent just like any other increment or decrement to lifetime resources: if less is spent at first (because $t_1 < T$), then more will be spent later. Thus if we want to inquire about the "effectiveness" of temporary taxes, we must specify a time horizon. Over a long enough run, they must be just as "effective" as permanent ones.

b) The so-called zero effect view—that consumers ignore the surtax and consume as if it never happened—does not represent the PIH at all. Instead, that theory says that consumers should spend precisely what they would on receipt of a windfall gain (or loss) of $t_1 z$. In the illustrative calculation, this turns out to be the "9 percent effect" view.

2. *Caveats and Imperfections*

There are several reasons why surtaxes may affect spending more strongly than indicated by pure theory. First, tax-induced income changes that are not consumed must be saved. If windfall gains are used to purchase durable goods, consumer spending may rise much more strongly than consumption; the converse may happen when there are windfall losses. The magnitude of the marginal propensity to spend windfalls on durable goods is, of course, an empirical question.[3]

Second, some households may be subject to liquidity constraints that are usually ignored by the PIH. If we stay within the certainty context, these constrained households will react strongly to even temporary income changes.[4] Thus the aggregate MPC for a temporary tax is a weighted average of the low MPCs of unconstrained households and the high MPCs of constrained ones. Again, the importance of this phenomenon is an empirical question.

[2] By way of comparison, as $t_1 \to T$, the MPC $\to 5/6$.

[3] See, e.g., Darby (1972).

[4] See Blinder (1976) and Dolde (1978). For a look at one particular type of uncertainty, see Foley and Hellwig (1975), which shows that this result may not carry through to the uncertainty case.

Third, as Okun (1971) pointed out, consumer behavior depends on what people believe rather than on what the government announces. If consumers disbelieve the government when it tells them that a tax hike is only temporary, then the spending response will be greater than that suggested by a naive application of the PIH.[5] Since the perceived duration of the surtax, not the declared duration, is relevant from the standpoint of the PIH, this too raises an empirical issue.

Finally, we must recognize the possibility that households may not do the kind of rational long-term planning envisioned by Modigliani and Brumberg (1954) and Friedman (1957) or, what amounts to the same thing, have very high subjective discount rates. If they are very shortsighted, then temporary fluctuations in disposable income may have substantial effects on spending.

II. Previous Empirical Work

Okun's (1971) study opened the empirical debate on this issue. Using the consumption equations of four econometric models, he compared the "full effect" view that the 1968 surtax was just as effective as a permanent tax increase to the "zero effect" view that consumers totally ignored the surtax. While he concluded that the full effect view fit the data better, an intermediate "50 percent effect" view actually does better than either extreme.[6] Springer (1975) criticized Okun's econometric procedures and then performed a similar experiment with a consumption function based on the PIH. He concluded that the zero effect view performed better.

Juster (1977), using a series of savings equations based on the Houthakker-Taylor (1966) model, reached conclusions about the 1975 rebate similar to those of Okun for the 1968 surtax. But Modigliani and Steindel (1977) found that the rebate had very little impact over a horizon of 1 or 2 quarters. Their modified version of the life-cycle model implied, however, a virtually full effect view over a 6-quarter horizon. Modigliani and Steindel assumed that the nonrebate portions of the 1975 tax cuts were treated like permanent taxes and handled the 1968 surtax with dummy variables.

The existing empirical literature thus offers little consensus. The issue seems quite open.

III. The Distributed Lag Model of Consumption

The basic vehicle for investigating the effectiveness of temporary income taxes in this paper is the distributed lag version of the PIH.

[5] This point has rather less cogency with respect to tax cuts; but here too consumers may believe them to be more permanent than the government announces.

[6] On this, see Blinder and Solow (1974, pp. 107–9).

While this is the standard way of implementing the PIH empirically, the recent literature on rational expectations has seemed to raise doubts about its validity.[7] This section shows how the PIH and rational expectations together lead to an estimating equation very much like the one I use in this paper.

As Muth (1960) pointed out, the PIH is basically a *forward-looking* model of consumer behavior. It states that consumers, in deciding on their current spending, weigh their current asset holdings, their current income from labor, and their expected future income from labor. Specifically, permanent income at time t is defined as

$$Y_t^p = A_t + \sum_{s=0}^{\infty} \frac{{}_tY_{t+s}}{(1+r)^s}, \tag{1}$$

where A_t is the stock of real assets at the beginning of period t, Y_t is noninterest income in period t, and ${}_tY_{t+s}$ is the mathematical (i.e., rational) expectation of Y_{t+s} that is formed at time t. (By convention, ${}_tY_t = Y_t$.) However, if the stochastic process generating income can be described by a time-series model such as

$$Y_t = a_1Y_{t-1} + a_2Y_{t-2} + \ldots + a_{n+1}Y_{t-n-1} + \epsilon_t, \tag{2}$$

where ϵ_t is a white noise error term, then the resulting empirical model of consumption will be *backward looking*. For example, if the theoretical consumption function is

$$C_t = \delta + kY_t^p + u_t, \tag{3}$$

then the empirical consumption function will be

$$C_t = \delta + kA_t + b_0Y_t + b_1Y_{t-1} + \ldots + b_nY_{t-n} + u_t. \tag{4}$$

To see this, it is only necessary to note that the (rational) expectation of income in period $t + s$ can, in view of (2), be based only on the information set $\{Y_t, Y_{t-1}, Y_{t-2}, \ldots\}$. Thus, for example,

$$\begin{aligned} {}_tY_{t+1} &= a_1Y_t + a_2Y_{t-1} + a_3Y_{t-2} + \ldots + a_{n+1}Y_{t-n}, \\ {}_tY_{t+2} &= a_{1t}Y_{t+1} + a_2Y_t + a_3Y_{t-1} + \ldots + a_{n+1}Y_{t-n+1}, \\ &= (a_1^2 + a_2)Y_t + (a_1a_2 + a_3)Y_{t-1} + \ldots, \end{aligned}$$

and so on. Substituting all such expressions into the definition (1) and then into the consumption function (3), it is clear that (4) is derived. As pointed out by Sargent (1978) and others, the coefficients b_i in equation (4) will be complicated functions of the coefficients a_i in (2).

Suppose then that, as suggested by Dolde (1976), we can distinguish among two or more sources of income whose generating functions (2) may differ. The PIH in conjunction with rational expectations then

[7] See Lucas (1976) and esp. Hall (1978).

implies that the b's in (4) should follow a different pattern for each income source. A simple example will illustrate this point and also give us some feeling for possible magnitudes. Consider several income sources, each of which is generated by a first-order autoregressive:

$$Y_{it} = \rho_i Y_{i,t-1} + \epsilon_{it}. \tag{5}$$

Working out the expectations and plugging into (1) gives us a simple expression for the permanent income attributable to each source:

$$Y_{it}^p = \frac{1 + r}{1 + r - \rho_i} Y_{it}, \qquad \text{if } \rho < 1 + r. \tag{6}$$

Notice that, despite the long time horizon contemplated by the PIH, consumption depends only on current income. In general, consumption will depend on past income only up to lag n, where $n + 1$ is the longest lag considered in equation (2).

Now compare two income sources, one of which is entirely permanent ($\rho = 1$) and the other of which is entirely transitory ($\rho = 0$). A \$1.00 increase in the permanent component will, according to (6), raise permanent income by \$$(1 + r)/r$ and thus raise consumption by \$$k(1 + r)/r$. This may imply a very large immediate spending response.[8] By contrast, a \$1.00 fluctuation in the purely transitory component will, again according to (6), raise permanent income by only \$1.00 and thus raise consumption by only \$$k$. The lesson, of course, generalizes and applies far beyond the confines of first-order autoregressives: *income sources deemed to be more permanent will elicit prompter spending responses than income sources deemed to be more temporary.* The application of this principle to permanent versus temporary changes in taxes is apparent and immediate and was elucidated clearly by Lucas (1976). It is the basic notion underlying the empirical model to be developed in the next section.

However, lest confusion arise, I should stress that there is no sense in which the rationality of expectations is either assumed or imposed in the consumption functions estimated here. My point is only that the distributed lag formulation of the PIH is consistent with rational expectations. As Sargent (1978) has emphasized, rational expectations delivers a set of restrictions across equations (2) and (4) that can be imposed in estimating the two jointly. I have made no attempt to impose these restrictions here because my interest was in getting the best possible consumption-function estimates, not in testing rationality. Furthermore, it is well known that quite different models

[8] Suppose the rate of subjective time discounting is equal to the rate of interest, so that a constant consumption stream is optimal, and that B is the lifetime propensity to consume (i.e., $1 - B$ is the propensity to bequeath). Then k will be $Br/(1 + r)$, so that $k(1 + r)/r$ will be B, which is close to unity.

of consumption behavior (e.g., habit persistence) can lead to an estimating equation very much like (4). It is not my purpose to discriminate among alternative ways of arriving at (4).

IV. Derivation of an Estimating Equation

The preceding discussion makes it clear that different distributed lag coefficients might be associated with different sources of income. While the actual empirical analysis considered four types of income, the model is most readily explained if I suppose there are only two: income (positive or negative) attributable to temporary tax measures, which I denote as S_t ("special income"); and all other disposable income ("regular income"), which I denote as R_t. The R_t should not be confused with permanent income, since it has both permanent and transitory components. The basic idea underlying the estimating equation is that S_t is identifiably "less permanent" than R_t.

Suppose consumption responds to R_t according to a set of distributed lag weights: $w_j = \partial C_t / \partial R_{t-j}, j = 0, 1, \ldots, n$. Since the w_j depend on the stochastic process generating R_t, it is worth reporting that the deviations of R_t from a logarithmic time trend are well described by the following second-order autoregressive:[9]

$$\underset{}{y_t = \underset{(.09)}{1.28 y_{t-1}} - \underset{(.09)}{.35 y_{t-2}}}, \qquad R^2 = .91,\ \text{D-W} = 1.86.$$

When income follows a second-order autoregressive, permanent income as defined in (1) is:

$$Y_t^p = A_t + K_t + \frac{(1+r)^2}{(1+r)(1+r-a_1) - a_2} y_t + \frac{a_2(1+r)}{(1+r)(1+r-a_1) - a_2} y_{t-1},$$

if $a_1 + a_2/(1 + r) < 1 + r$, where K_t is the present value of the trend component of labor income and y_t and y_{t-1} are current and lagged deviations from trend. Given the estimates of a_1 and a_2 above, and for $r = .0074$ (a 3 percent annual real interest rate), the implied coefficients are $Y_t^p = A_t + K_t + 13.5y_t - 4.7y_{t-1}$. This leads us to expect a very large value of w_0, followed by swiftly declining w's—possibly even turning negative. The empirical results bear this out.

As Lucas (1976) has argued, income changes that are clearly "more temporary" than regular income should get different spending

[9] Standard errors are in parentheses. Longer autoregressives, however, give slightly better fits. E.g., if $t - 11$ is the longest lag allowed to enter the regression, significant coefficients are obtained at lags 1, 2, 3, 4, and 11. An F-test for the zero restrictions implied by the second-order model, however, yields an F-ratio of only 1.77, which is well below the critical 5 percent point of the χ_9^2 distribution.

coefficients. To develop a model of the distributed lag response of C_t to S_t, first break down S_t into its components:

$$S_t = S_t^1 + S_t^2 + \ldots + S_t^m, \tag{7}$$

where S_t^i indicates the income gain or loss in quarter t from the ith temporary tax. (In the empirical work, $m = 3$.) It will help clarify the treatment of the S_t^i if I define a hypothetical set of lag coefficients β_j as the effect on C_t of a \$1.00 pure windfall gain received in quarter $t - j$.

The treatment of S_t^i depends on whether or not the ith temporary tax is still in effect. If it is, I assume that S_t^i is treated as a weighted average of regular and windfall income, so that it gets the distributed lag weights

$$\gamma_j = \frac{\partial C_t}{\partial S_{t-j}^i} = \lambda w_j + (1 - \lambda)\beta_j, \qquad \begin{array}{l} j = 0, 1, \ldots, n \\ 0 \leq \lambda \leq 1. \end{array} \tag{8}$$

If the temporary tax is no longer on the books, I assume that consumers look upon S_{t-j}^i in retrospect as if it had been a pure windfall and so apply the distributed lag coefficients β_j. By introducing a dummy variable defined as:

$$D_t^i = 1 \text{ if the } i\text{th temporary tax remains in force in quarter } t,$$
$$= 0 \text{ otherwise,}$$

it is possible to combine these two hypotheses into a single expression:

$$\gamma_j^i(t) = D_t^i[\lambda w_j + (1 - \lambda)\beta_j] + (1 - D_t^i)\beta_j, \tag{9}$$

where the notation now indicates that the γ weights depend both on calendar time and on the specific tax under consideration (because of the dummy variable).

An interesting point arises here. Standard pre-rational-expectations approaches to consumption-function estimation would suggest that $\Sigma\beta_j$ and $\Sigma\gamma_j$ be constrained to equal Σw_j, apparently meaning that the "long-run MPC" out of any type of income is identical. However, the PIH-cum-rational-expectations approach suggests no such adding-up constraint. To see this, follow Sargent (1978, pp. 681–82) in rewriting (2) in the form $X_t = HX_{t-1} + \eta_t$ (Sargent's eq. 8), where

$$X_t = \begin{bmatrix} Y_t \\ Y_{t-1} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ Y_{t-n} \end{bmatrix}, \quad H = \begin{bmatrix} a_1 & a_2 & \cdots & a_n & a_{n+1} \\ 1 & 0 & & 0 & 0 \\ 0 & 1 & & \cdot & \cdot \\ \cdot & 0 & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & 0 & \cdot \\ 0 & 0 & & 1 & 0 \end{bmatrix}, \quad \eta_t = \begin{bmatrix} \epsilon_t \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix},$$

and $Y_t = dX_t$, where $d = (1, 0, \ldots, 0)$. As Sargent notes (his eq. 9), rational expectations implies ${}_tX_{t+s} = H^sX_t$, whence ${}_tY_{t+s} = dH^sX_t$. Substituting this into (1) gives the following expression for permanent income:

$$Y_t^p = A_t + \left[\sum_{s=0}^{\infty} \frac{dH^s}{(1+r)^s}\right] X_t,$$

which is of the form (4) with coefficients

$$b \equiv (b_0, b_1, \ldots, b_n) = \sum_{s=0}^{\infty} \frac{dH^s}{(1+r)^s}.$$

The sum of the b_j has no obvious interpretation. Thus, in a model with several sources of income, there is no particular reason why the various sets of distributed lag coefficients should have a common sum.

Where, then, does the lifetime budget constraint enter? The answer is that (4) implies a unitary lifetime MPC for any values of k and the b_j. The proof involves some straightforward but tedious algebraic manipulations of the difference equations (4) and

$$A_{t+1} = (1 + r)A_t + Y_t - C_t \tag{10}$$

and hence is relegated to Appendix A.

With these preliminaries out of the way, it is easy to explain the estimating equation. If there were no special taxes to worry about, the basic empirical model of consumer behavior would be as follows:

$$\begin{aligned} C_t = k_0 + k_1 r_t Y_t + \sum_{j=0}^{n} w_j Y_{t-j} \\ + k_2(W_t - A_t) + \sum_{j=0}^{q} k_{3+j} A_{t-j} + u_t, \end{aligned} \tag{11}$$

where r_t is the rate of interest, W_t is consumer net worth at the beginning of period t, and A_t is the market value of stock market wealth at the beginning of period t. The specific way in which assets are entered into the consumption function, including the constraint that $k_3 + k_4 + \ldots + k_{3+q} = k_2$, is suggested by the MIT-Penn-SSRC (MPS) model and is unimportant to what follows.

Now consider the separation of disposable income into its two components:

$$Y_t = R_t + S_t. \tag{12}$$

The way I have defined the γ's means that (11) is expanded to:

$$C_t = k_0 + k_1 r_t Y_t + k_2(W_t - A_t) + \sum_{j=0}^{q} k_{3+j} A_{t-j}$$

$$+ \sum_{j=0}^{n} w_j R_{t-j} + \sum_{j=0}^{n} \gamma_j^1(t) S_{t-j}^1 \tag{13}$$
$$+ \dots + \sum_{j=0}^{n} \gamma_j^m(t) S_{t-j}^m + u_t.$$

Substituting (9) and (12) into (13), and rearranging terms, gives:

$$\begin{aligned} C_t = k_0 + k_1 r_t Y_t + k_2(W_t - A_t) + \sum_{j=0}^{q} k_{3+j} A_{t-j} \\ + \sum_{j=0}^{n} w_j (Y_{t-j} + \lambda X_t^j - S_{t-j}) \\ + \sum_{j=0}^{n} \beta_j \left[(1-\lambda) X_t^j + \sum_{i=1}^{m} (1 - D_t^i) S_{t-j}^i \right] + u_t, \end{aligned} \tag{14}$$

where $X_t^j \equiv D_t^1 S_{t-j}^1 + \dots + D_t^m S_{t-j}^m, j = 0, \dots, n$. This is not the actual estimating equation because additional income sources were distinguished, because the distributed lag coefficients were constrained in several ways, and because corrections were made for both heteroscedasticity and serial correlation in the error term. Details are spelled out in Appendix B. Nonetheless (14) is the most useful form for interpreting the estimated parameters. The model is nonlinear because of the parameter λ—the crucial parameter of this study.

As noted above, theory does not imply that $\Sigma w_j = \Sigma \beta_j$. To investigate this further, this adding-up constraint was imposed in (14) and its validity tested as follows.[10] Let ℓ denote the likelihood ratio. Then, under the assumption of normality, $-2 \log \ell = T \log (\mathrm{SSR}_r/\mathrm{SSR})$ is distributed as a χ^2 with r degrees of freedom, where: T = number of observations (= 100 throughout this paper), SSR = minimized value of the sum of squared residuals in the unconstrained regression (eq. [14] in this case), SSR_r = minimized value of the sum of squared residuals in the constrained regression (obtained by imposing $\Sigma\beta = \Sigma w$), and r = number of restrictions (= 1 in this case). As reported in table 1, row 1, the constraint was rejected at the 10 percent level but not at the 5 percent level. There being no persuasive theoretical rationale for it, the constraint was dropped.

Tacitly, however, (14) embodies a number of other constraints that are equally lacking in theoretical justification—constraints that each type of regular income is subject to the same set of distributed lag coefficients. These constraints were tested by a series of likelihood ratio tests, which are described in the balance of this section.

[10] See Goldfeld and Quandt (1972, p. 74).

TABLE 1

χ^2 TESTS OF CONSTRAINTS

Unconstrained Model	Constraint Tested	df	Test Statistic
1. Eq. (14)	$\Sigma w = \Sigma \beta$	1	3.67
2. Eq. (16)	$v_j = w_j$ for all j	3	12.02
3. Eq. (16)	$\Sigma v_j = \Sigma w_j$	1	Approximately 0*
4. Eq. (18) with $\Sigma v = \Sigma w$	$\phi_j = w_j$ for all j	3	19.02
5. Eq. (18) with $\Sigma v = \Sigma w$	$\Sigma \phi_j = \Sigma w_j$	1	12.52
6. Eq. (18) with $\Sigma v = \Sigma w$	$\lambda = 0$	1	1.32
7. Eq. (18) with $\Sigma v = \Sigma w$	$\lambda = 1$	1	2.84
8. Eq. (18) with $\Sigma v = \Sigma w$	$k_2 = 0$	1	17.81

NOTE.—Critical levels for the χ^2 distribution are:

df	10% Point	5% Point	1% Point
1	2.71	3.84	6.63
3	6.25	7.81	11.34

*Due to rounding error, the actual computed test statistic was slightly negative. When this constraint was tested in the context of eq. (18), it produced a test statistic of 1.32.

First, "regular" disposable income was disaggregated into its two main components—personal income and "regular" personal taxes:

$$R_t = P_t - T_t.^{11} \tag{15}$$

Personal income is assumed to be spent according to the lag weights w_j, while regular taxes are assumed to be subject to a different set of lag weights v_j. Special taxes are treated as previously explained, except that the v's replace the w's in equations (8) and (9). That is, while the tax is on the books, a special tax is treated as a weighted average of a permanent tax and a windfall. Thus the basic consumption function becomes:

$$\begin{aligned} C_t = {} & k_0 + k_1 r_t Y_t + k_2(W_t - A_t) + \sum_{j=0}^{q} k_{3+j} A_{t-j} \\ & + \sum_{j=0}^{n} w_j P_{t-j} - \sum_{j=0}^{n} v_j (T_{t-j} - \lambda X_t^j) \\ & + \sum_{j=0}^{n} \beta_j \left[(1-\lambda) X_t^j + \sum_{i=1}^{m} (1 - D_t^i) S_{t-j}^i \right] + u_t. \end{aligned} \tag{16}$$

Since (as explained below) the distributed lag coefficients are constrained to follow a third-degree polynomial with a zero end-point

[11] In making this separation, I departed a bit from national income accounting conventions by including the employer's share of social insurance contributions in both P and T.

constraint, the null hypothesis that the v_j are equal to the w_j imposes three constraints on equation (16). Row 2 of table 1 shows that these constraints were resoundingly rejected by the data ($\chi^2_3 = 12$). However, it turned out that the sum of the v_j was estimated to be almost exactly equal to the sum of the w_j (see row 3 of table 1), so this adding-up constraint was imposed in subsequent estimates.

The final generalization considered was to disaggregate personal income into its two main components—factor income and transfer payments:[12]

$$P_t = F_t + V_t. \tag{17}$$

This made the estimating equation:

$$\begin{aligned} C_t = k_0 + k_1 r_t Y_t + k_2(W_t - A_t) + \sum_{j=0}^{q} k_{3+j} A_{t-j} \\ + \sum_{j=0}^{n} w_j F_{t-j} + \sum_{j=0}^{n} \phi_j V_{t-j} - \sum_{j=0}^{n} v_j (T_{t-j} - \lambda X_t^j) \\ + \sum_{j=0}^{n} \beta_j \Big[(1 - \lambda) X_t^j + \sum_{i=1}^{m} (1 - D_t^i) S_{t-j}^i\Big] + u_t. \end{aligned} \tag{18}$$

Once again, the null hypothesis that the ϕ_j (spending coefficients for transfers) are in fact equal to the w_j (spending coefficients for factor income) was tested by a χ^2 test. And once again it was resoundingly rejected ($\chi^2_3 = 19$; see row 4 of table 1). This time, however, the data also rejected the adding-up constraint $\Sigma\phi_j = \Sigma w_j$, which therefore was not imposed (table 1, row 5). In fact, most of the difference between the ϕ_j and the w_j was in their sums; the time patterns were remarkably similar.

To summarize these tests, we are left with a model that assigns distributed lag weights w_j to factor income, ϕ_j to transfers, v_j to regular taxes, and β_j to windfalls. The Σv_j and Σw_j are apparently equal, but the other sums are not.

V. Issues in Estimation

1. *Data*

Following the suggestion of Darby (1975), I used consumer expenditures, rather than pure consumption, as C_t. This seems most appropriate where the focus is on the evaluation of stabilization policy, as it is here, rather than on testing the PIH. Furthermore, it

[12] For this purpose, both employer contributions and business transfers are considered to be factor income, and the aspects of the 1975–76 tax cuts that are classified as transfer payments in the national income accounts are grouped with temporary taxes.

TABLE 2

EFFECTS ON DISPOSABLE INCOME OF 1975–76 TAX CUTS AND TRANSFERS*

	Taxes†		Transfers‡		
Quarter	Tax Rate Cuts	Rebate	Social Security	Earned Income	Total
1975:2	8.5	31.2	6.7	0	46.4
1975:3	12.3	0	0	0	12.3
1975:4	11.9	0	0	0	11.9
1976:1	14.2	0	0	1.9	16.1
1976:2	14.0	0	0	1.6	15.6

* In billions of current dollars, at annual rates.
† From U.S. Bureau of Economic Analysis (February 1976, March 1977).
‡ Kindly supplied to the author by Joseph C. Wakefield, of the Bureau of Economic Analysis, in conversation.

avoids many complicated issues of definition (e.g., Which goods are durables? How fast do they depreciate? etc., etc.). The cost of this shortcut is that the theoretical interpretation of some of the parameters is lost. For example, k_1 includes the effects of r_t on both pure consumption (which may be positive or negative) and spending on durables (which should be negative). Similarly, the lag weights w_j might be expected to be less smooth than the lag of pure consumption behind income because of the lumpy nature of durables.[13]

Data for the 1968 surcharge were taken from Okun (1971) and converted to 1972 dollars by the deflator for personal consumer expenditures; these comprise S_t^1. Data for the various 1975–76 tax cuts are shown in table 2; they were similarly deflated and segregated into two time series. The S_t^2 series is defined as the explicitly one-shot measures: the tax rebates and the social security bonuses (hereafter referred to as "the rebate"). The rest is considered as S_t^3. Since the 1975 cuts were extended several times and are now a permanent feature of the tax code, an arbitrary decision had to be made as to when they became "permanent."[14] I decided to cut off S_t^3 after 1976:2, because by then the cuts had already been extended once in the Revenue Adjustment Act of 1975 and a second time in the Tax Reform Act of 1976.

[13] To account for the special features of expenditures on consumer durables, several additional variables were tried in some earlier regressions. Neither the stock of durables, nor the relative price of durables, nor the unemployment rate succeeded in significantly lowering the sum of squared residuals. In several cases, the signs of the coefficients were even the opposite of what theory suggests. It may be that these variables are more relevant to the choice between saving in the form of durables vs. in financial form than they are to the choice between spending and saving.

[14] At an early stage of this research, I experimented with a learning model in which a temporary tax gradually came to be considered permanent as it remained on the books longer and longer. This experiment was unsuccessful.

Data on consumer net worth (W_t), and its breakdown into stock market (A_t) and non–stock market ($W_t - A_t$) components, were taken from the data bank of the MPS model and converted to 1972 dollars. They are based on a number of primary sources, the most important of which is the flow of funds.

Because of the recent findings of Boskin (1978), I thought it important to use a real after-tax interest rate. Since the construction of this series was a fairly involved affair, I explain it only in Appendix C.

2. *Distributed Lag Estimation*

The many distributed lags in equation (18) were estimated by an adaptation of the Almon (1965) lag technique, as a method of conserving on parameters. Generally, a third-degree polynomial with a zero constraint at the far end was used. Preliminary tests suggested that these end-point constraints ($w_{n+1} = 0$, etc.) could not be rejected. An unconstrained version of one specification, run as an experiment, showed that the polynomial constraint had very little effect on the w_j coefficients but did influence the β_j coefficients.

The distributed lag effects of assets were estimated as follows. In some preliminary regressions, the two components were combined and a cubic distributed lag over n quarters was estimated. Then the two components were disaggregated. These preliminary tests showed quite clearly (a) that the coefficients were not the same and (b) that the lag was much shorter than n quarters. (As explained just below, n was chosen to be 7.) In the case of non–stock market wealth, an estimated distributed lag over 4 quarters attached virtually all of the weight to the current (start of period) value, so all lagged values were omitted. In the case of stock market wealth, when a cubic was estimated over $j = 0, 1, \ldots, 7$, the coefficients turned out to be almost linear and to be virtually zero after $j = 2$ (a small positive value for $j = 3$ and small negative values for $j = 4, \ldots, 7$). Thus a linear distributed lag over $j = 0, \ldots, 3$ was adopted as the final specification.

The length of the distributed lag, n, was selected by running a preliminary version of the regression for alternative values of n ranging from 6 to 10. A very clear minimum in the sum of squared residuals was found around $n = 7$ or $n = 8$, with the former having a slightly better fit and slightly better coefficients. On this basis, $n = 7$ was selected for all subsequent work.

3. *Treatment of the 1975 Rebate*

As has been noted already, the model divides the 1975–76 tax cuts into two parts: S_t^2 includes the rebate, while S_t^3 includes all the rest.

This makes a strong (and questionable) assumption about how consumers treated the rebate. In particular, it assumes that they treated it just like the 1968 surcharge and the other 1975 reductions: essentially as if a fraction λ of it was a regular increase in income, while a fraction $1 - \lambda$ was a pure windfall. Given the nature of the rebate, this is questionable, to say the least.

An alternative assumption—equally strong as the first—is that consumers treated the rebate as a pure windfall right from the start.[15] While it is not obvious that this is true, since consumers might have anticipated a repeat performance with some reasonable probability, it does seem a plausible working hypothesis. Fortunately, it is not difficult to modify any of the three models to accommodate this alternative hypothesis; all that is necessary is that the lag weights β_j be applied to S_t^2 starting immediately in 1975:2. When this was done in one version of the model, the resulting equation had virtually an identical SSR, almost the same estimated λ, and very similar implications about spending patterns out of the rebate. Thus the conclusions of this study seem insensitive to the treatment of the rebate.

VI. Empirical Results

1. *Parameter Estimates and Interpretation*

Estimation was done by the numerical optimization package developed by S. M. Goldfeld and R. E. Quandt. The results from estimating equation (18) on quarterly data covering 1953:1–1977:4 are presented in table 3. The number in parentheses next to each estimated coefficient is its asymptotic standard error (or rather a numerical estimate thereof).

In interpreting the standard error of the regression, it should be mentioned that, as explained in Appendix B, the equation was actually transformed so that the left-hand variable was the average propensity to consume (APC), C_t/Y_t, rather than consumer spending. Thus, the standard error of .0038 is relative to a typical value for the APC of about .90. This represents an excellent fit.[16] At 1977 income levels, it translates to a standard error of about \$3.5 billion in predicting consumer spending. Of course, obtaining a good fit with a consumption function is hardly a notable achievement, and the equation—like most consumption functions—does suffer from some autocorrelation ($\rho = .54$).

[15] Since the rebate was off the books by 1975:3, and hence treated as a windfall in any case, only 1975:2 is at issue here.

[16] The standard errors of a comparable equation in Modigliani and Steindel (1977) are .0056 with an autocorrelation correction and .0065 without.

TABLE 3

NONLINEAR CONSUMPTION FUNCTION PARAMETER ESTIMATES*

k_0	k_1	k_2	λ	ρ
30.1 (7.1)	.0002 (.0007)	.021 (.005)	.50 (.32)	.54 (.11)

DISTRIBUTED LAG COEFFICIENTS†

j	w_j	ϕ_j	v_j	β_j	k_{3+j}	γ_j	Rebate
0	.60 (.05)	.50 (.16)	.34 (.09)	−.03 (.26)	.009 (.002)	.16	.16
1	.16 (.02)	−.02 (.09)	.14 (.04)	−.03 (.11)	.006 (.001)	.06	−.03
2	−.06 (.03)	−.20 (.10)	.04 (.05)	−.01 (.09)	.004 (.001)	.02	−.01
3	−.12 (.02)	−.14 (.08)	.01 (.05)	.03 (.09)	.002 (.0005)	.02	.03
4	−.08 (.01)	.04 (.06)	.02 (.03)	.08 (.07)	...	.05	.08
5	.02 (.01)	.26 (.07)	.05 (.03)	.11 (.06)	...	.08	.11
6	.10 (.02)	.39 (.09)	.08 (.04)	.12 (.08)	...	.10	.12
7	.11 (.02)	.34 (.08)	.07 (.04)	.09 (.07)	...	.08	.09
Sum	.74	1.17	.74	.36	.021 (.005)	.55	.55

NOTE.—k_0 = constant, k_1 = coefficient of interest rate, k_2 = coefficient of non–stock market wealth, λ = weight on regular income, ρ = autocorrelation coefficient. Sum of squared residuals = .00147, SE = .00383, SE of unadjusted errors (without a correction for autocorrelation) = .00455, N observations = 100.

* Asymptotic SEs are in parentheses.

† Components may not add to totals due to rounding.

Turning to the coefficient estimates, the most critical parameter for purposes of this study is λ, the weight attached to regular income in equation (8). The point estimate of .50 suggests that temporary taxes that are still on the books are treated like 50-50 blends of windfalls and regular taxes. However, the standard error is regrettably large; there are, after all, pitifully few observations that can be used to estimate λ. The null hypotheses $\lambda = 0$ or $\lambda = 1$ can nonetheless be tested by likelihood ratio tests. When these tests were run with equation (18) as the unconstrained regression (see table 1, rows 6 and 7), the null hypothesis that $\lambda = 0$ (temporary taxes are regarded as pure windfalls) could not be rejected ($\chi_1^2 = 1.32$). But the null hypothesis that $\lambda = 1$ (temporary taxes are regarded as regular income) could be rejected if we were not too fussy about significance levels ($\chi_1^2 = 2.84$).

I turn next to the distributed lag coefficients of the various income terms. The w's for factor income are very large and positive at first, then turn small and negative, and finally become positive again at the end. This general shape accords well with our expectations.[17] The ϕ_j coefficients for transfer payments follow a similar shape but are much more erratic and less well pinned down econometrically. A notable

[17] Because of the estimating form, it is unlikely that simultaneity has much to do with the large estimate for w_0. See Appendix B.

feature is that their sum is nearly 1.2, indicating "overspending" during the first 2 years after receipt of a transfer payment.

The v_j coefficients for regular taxes also exhibit a characteristic U-shape, but in much more muted fashion. As compared with factor income, spending in the first year after a regular tax cut is apparently substantially less, after which it catches up. This was surprising at first, since Dolde (1976) and Modigliani and Steindel (1977) had suggested that regular tax changes are "more permanent" than regular income.[18] However, it turns out that the following second-order autoregressives describe the deviations from trend of personal income (P_t) and regular taxes (T_t):

$$P_t = \underset{(.09)}{1.32}P_{t-1} - \underset{(.09)}{.38}P_{t-2},$$

$$T_t = \underset{(.09)}{1.11}T_{t-1} - \underset{(.10)}{.21}T_{t-2}.$$

Using the formulas derived earlier for permanent income, these time-series models imply that "permanent personal income" and "permanent taxes" are given by $P_t^p = 14.9P_t - 5.7P_{t-1}$ and $T_t^p = 9.3T_t - 1.9T_{t-1}$, so that, contrary to Dolde and Modigliani and Steindel, we should actually expect a stronger short-run response of consumption to fluctuations in P_t than to fluctuations in T_t, which is exactly what I find.

The next column, the β_j's, are in some sense out-of-sample extrapolations since there are no "pure windfalls" recorded in the data.[19] Their only use is to form the weighted average $\lambda w_j + (1 - \lambda)\beta_j$, which is reported in the column marked "γ_j." These are the expenditure coefficients for income from a temporary tax that remains on the books for the entire 2-year horizon. To illustrate the opposite extreme, the column marked "Rebate" shows the spending coefficients for a temporary tax that lasts only 1 quarter. These two columns differ in details but are quite similar. There is a moderate spending response in the initial quarter, followed by very little spending over the next 3 or 4 quarters. Most of the spending out of a temporary tax cut, according to these estimates, comes 5 or more quarters after the cut.

The coefficient of assets (.02) is comparable to what others have estimated, though a bit on the low side. A likelihood ratio test of the null hypothesis $k_2 = 0$ (which, in this constrained form, also implies k_3

[18] But see Dolde (1979), where the transitory nature of allegedly permanent tax changes is stressed.

[19] I made an attempt, in some early regressions, to treat the National Service Life Insurance Dividends of 1950 in this way, but I was not successful.

$= \ldots = k_6 = 0$) produced a test statistic of 17.8, which is highly significant at any reasonable significance level (table 1, row 8).

If we ignore the fact that income from property is included in the measure of income, the parameter k_2 can be given an interesting theoretical interpretation. In the basic life-cycle model, the consumer maximizes a utility function of the form

$$\sum_{t=0}^{T} \left(\frac{1}{1+\rho}\right)^t \frac{C_t^{1-\delta}}{1-\delta},$$

subject to a lifetime wealth constraint. It can be shown by straightforward computations that the optimal solution for initial consumption is $C_0 = [(r-g)/(1+r)]W$, where W is lifetime wealth and g is the optimal growth rate of C_t, defined as $g = [(1+r)/(1+\rho)]^{1/\delta} - 1$. This means that k_2 corresponds to the theoretical coefficient $(r-g)/(1+r)$, which for small values of r and ρ is approximately equal to $r + (1/\delta)(\rho - r)$. Thus for small values of r the estimated value $k_2 = .02$ implies that the subjective discount rate, ρ, is approximately $.02\delta$ or around 2–3 percent per quarter for plausible values of δ.

One striking result, though it is peripheral to the subject of this study, is the tiny coefficient of the real after-tax interest rate.[20] This finding turned up in every specification of the model, including several alternative measures of the rate of interest. (Sometimes the coefficient was trivially negative, sometimes trivially positive, but always trivial.) While it accords well both with my earlier work (Blinder 1975) and with the work of others, it stands in sharp contrast with Boskin's (1978) recent finding of a strong positive interest elasticity of saving.

2. *Temporary versus Permanent Taxes*

We can now address the principal issue of this study: How effective are explicitly temporary income tax changes as compared to those announced to be permanent? Table 4 contains the answers derived from the model, using the parameter estimates presented in table 3 to make equation (4) operational and using an annual real interest rate of 3 percent in updating wealth according to equation (10). It can be seen from column 4 that a temporary tax is about one-half as effective as a permanent tax in the first year, rising to about three-quarters as effective in the second year. Spending out of a rebate is somewhat slower than this. My estimated cumulative spending propensities out

[20] The specific coefficient in table 3 means that a 1 percentage point rise in r lowers savings by about .02 of 1 percent of disposable income—a trivial amount.

TABLE 4

RELATIVE EFFECTIVENESS OF TEMPORARY TAXES

	CUMULATIVE SPENDING PROPENSITIES			RATIOS	
j	Permanent (1)	2-Year (2)	Rebate (3)	(2)/(1) (4)	(3)/(1) (5)
0	.34	.16	.16	.47	.47
1	.50	.23	.14	.46	.28
2	.55	.26	.16	.47	.29
3	.56	.30	.21	.54	.38
4	.59	.36	.30	.61	.51
5	.65	.46	.43	.71	.66
6	.73	.57	.56	.78	.77
7	.81	.65	.66	.80	.81

of a rebate are larger than those estimated by Modigliani and Steindel (1977) for the first few quarters but smaller thereafter.

These findings carry two important messages to fiscal policy planners. First, and most obvious, is that temporary taxes are less powerful devices for short-run stabilization purposes than are permanent ones. Second, and perhaps almost as important, the short-run relative ineffectiveness of such taxes implies that the impact of these measures in the second year is larger than might be expected. For example, according to table 4, each \$1.00 of permanent tax reduction adds \$0.25 to spending in the second year, while each \$1.00 of a rebate adds \$0.45. If the need is for a truly short-run stimulus to aggregate demand, this effect may also be unwanted.

Both of these points can be illustrated by examining what the equations have to say about the 1975–76 episode. First, it is useful to display the observed APCs for this period in table 5. There are two obvious phenomena crying out for explanation in these data. First, why did the APC drop so sharply in 1975:2? Second, why did it thereafter begin a steady climb to what is a truly extraordinary level by 1977:1? (The corresponding personal savings rate was only 4.2

TABLE 5

AVERAGE PROPENSITIES TO CONSUME, 1975–77

	1975	1976	1977
First quarter	.913	.914	.935
Second quarter	.881	.918	.926
Third quarter	.903	.922	.922
Fourth quarter	.907	.926	.925

SOURCE.—U.S. Bureau of Economic Analysis (various issues).

TABLE 6

Estimated Effects of the 1975–76 Temporary Tax Cuts on Consumer Expenditures*

Quarter	Estimated Spending Effect
1975:2	5.9
1975:3	1.6
1975:4	2.7
1976:1	4.8
1976:2	6.7
1976:3	6.7
1976:4	7.1
1977:1	6.6
1977:2	7.2
1977:3	7.5
1977:4	7.3

* In billions of 1972 dollars.

percent—the lowest figure that had then been recorded since the Korean War.)

According to the estimates presented in this paper, the temporary tax cuts of 1975–76 contributed to both phenomena. Using the spending coefficients presented in table 4, table 6 shows the estimated direct (excluding multiplier) effects on consumer spending of the tax cuts of 1975:2 through 1976:2, inclusive.[21] It appears that (1) very little of the rebate was spent in 1975:2, (2) rather little of the disposable income attributable to the temporary tax cut package was spent during the remainder of 1975, and (3) more spending out of the temporary tax cuts was done in 1976 and yet more in 1977. Both the low APC of 1975:2 and the high APCs of late 1976 and early 1977 are tracked very well by the model. One observation which virtually jumps from table 6 is how very small these estimated spending impacts are relative to the size of the economy they were meant to stimulate (real GNP in the neighborhood of $1,250 billion).

Finally, there is one more question. If, instead of the 1975:2–1976:2 package of temporary measures, the government had cut taxes "permanently" in 1975:2 and then restored them to their original level starting in 1976:3, how large a tax cut would have achieved the same average effect on aggregate demand?[22]

Table 7 summarizes the model's answers to this question for three different choices of the horizon over which the "average effect on aggregate demand" might be defined. The first column gives the average direct impact on spending attributed by the model to the

[21] The reader is reminded that only these 5 quarters are considered temporary cuts.

[22] For this calculation I assume that consumers were successfully fooled into thinking that the 5-quarter tax cut would be permanent.

TABLE 7

EQUIVALENT PERMANENT TAXES*,†

Horizon	Average Impact on Spending of Actual 1975–76 Cuts	Cumulative Revenue Loss over 5 Quarters of Permanent Tax Cut with Equal Average Impact on Spending
4 quarters	3.7	9.5
6 quarters	4.7	12.4
8 quarters	5.3	15.9

* See text for definition.
† In billions of 1972 dollars, at annual rates.

1975–76 tax cuts. The next column shows how much total tax revenue the government would have had to relinquish during the same 5-quarter period (1975:2–1976:2) in order to achieve the same direct impact on spending through a permanent tax cut. Since the total 5-quarter revenue loss from the 1975–76 package was $20 billion, these numbers mean, for example, that a permanent tax cut about half as large ($9.5 billion vs. $20 billion) would have had the same first-year effect on aggregate demand. Over a 2-year horizon, however, the 1975–76 package had about 80 percent as much "bang for the buck" as a permanent tax cut.

VII. Summary

Both economic theory and casual empirical observation of the U.S. economy suggest that short-run spending propensities from temporary tax changes are smaller than those from permanent ones, but neither provides much guidance about the magnitude of this difference. This paper offers new empirical estimates of this difference and finds it to be quite substantial.

The analysis is based on an amendment of the standard distributed lag version of the PIH that distinguishes temporary taxes from other income on the grounds that the latter is "more transitory." This amendment, which is broadly consistent with rational expectations, leads to a nonlinear consumption function.

Though the standard error is unavoidably large, the point estimate suggests that a temporary tax change is treated as a 50-50 blend of a normal income tax change and a pure windfall. Over a 1-year planning horizon, a temporary tax change is estimated to have only a little more than half as much impact as a permanent tax change of equal magnitude, and a rebate is estimated to have only about 38 percent as much impact. The model tracks both the extraordinarily high savings rate of 1975:2 and the extraordinarily low savings rates of late 1976 and early 1977 very well and attributes part of both phenomena to the temporary tax measures of 1975–76. Finally, it is estimated that a

permanent tax cut of about \$9.5 billion (in 1972 dollars) would have had the same impact on aggregate demand over the first 4 quarters as the \$20 billion of 1975–76 tax cuts.

Appendix A

The Lifetime Budget Constraint

This Appendix demonstrates a result that, to my knowledge, is not very well known: that the long-run MPC corresponding to a consumption function of the form

$$C_t = kA_t + \sum_{j=0}^{n} b_j Y_{t-j} \tag{A1}$$

is unity for any value of k (greater than the real rate of interest) and for any b's.

Proof: Write equation (10) in the text as

$$[1 - (1 + r)L]A_t = Y_{t-1} - C_{t-1}, \tag{A2}$$

and write (A1) as

$$C_t = kA_t + b(L)Y_t, \tag{A3}$$

where L is the lag operator and $b(L) = b_0 + b_1L + \ldots + b_nL^n$. Applying the operator $1 - (1 + r)L$ to (A3) and using (A2) yields $[1 - (1 + r - k)L]C_t = \{kL + [1 - (1 + r)L]b(L)\}Y_t$, which can be written

$$C_t = B(L)Y_t, \tag{A4}$$

where

$$B(L) = \frac{kL + [1 - (1 + r)L]b(L)}{1 - (1 + r - k)L}. \tag{A5}$$

To obtain the lifetime spending generated by a \$1.00 impulse in Y_t, we must compute the discounted sum of coefficients:

$$\text{MPC} = \sum_{i=0}^{\infty} \frac{B_i}{(1 + r)^i}. \tag{A6}$$

To simplify the notation, let $\theta \equiv 1 + r - k$. Assuming that $\theta < 1$ (i.e., that $k > r$), (A5) can be written: $B(L) = \{kL + [1 - (1 + r)L]b(L)\}(1 + \theta L + \theta^2L^2 + \ldots)$. This is of the form $B(L) = B_0 + B_1L + B_2L^2 + \ldots$, with the following coefficients:

$$B_0 = b_0,$$

$$B_1 = \theta b_0 + [b_1 + k - (1 + r)b_0],$$

$$B_2 = \theta^2 b_0 + \theta[b_1 + k - (1 + r)b_0] + [b_2 - (1 + r)b_1],$$

$$\vdots$$

$$B_n = \theta^n b_0 + \theta^{n-1}[b_1 + k - (1 + r)b_0] + \ldots + [b_n - (1 + r)b_{n-1}],$$

$$B_{n+1} = \theta^{n+1} b_0 + \theta^n[b_1 + k - (1 + r)b_0] + \ldots + \theta[b_n - (1 + r)b_{n-1}] - (1 + r)b_n,$$

$$B_{n+1+s} = \theta^s B_{n+1}, \qquad s = 1, 2, \ldots .$$

Substitution of all of these into (A6), and some truly horrendous grinding, establishes that MPC = 1 regardless of the magnitudes of k and the b_j (as long as $k > 0$). Q.E.D.

A brief word on the interpretation of the sum of the b_j in (A1) may be in order here. Should Y_t rise permanently by $1.00—a statement that is basically meaningless if the autoregressive process assumed in the text (eq. [2]) really holds—the eventual change in C_t would, by (A4), be $B(1) = \Sigma_{i=0}^{\infty} B_i$. According to (A5), this sum is

$$B(1) = \frac{k - r\sum_{j=0} b_j}{k - r} = 1 + \frac{r}{k - r}\left(1 - \sum_{j=0}^{n} b_j\right).$$

Thus Σb_j controls the size of the spending response to a hypothetical permanent rise in income; it does not influence the lifetime MPC.

Appendix B

Details on the Estimating Equation

This Appendix derives and explains the equation that was actually estimated. I begin by repeating equation (14) of the text:

$$\begin{aligned} C_t &= k_0 + k_1 r_t Y_t + k_2(W_t - A_t) + \sum_{j=0}^{q} k_{3+j} A_{t-j} \\ &\quad + \sum_{j=0}^{n} w_j(Y_{t-j} + \lambda X_t^j - S_{t-j}) + \sum_{j=0}^{n} \beta_j \left[(1 - \lambda)X_t^j + \sum_{i=1}^{m} (1 - D_t^i)S_{t-j}^i\right] + u_t. \end{aligned}$$

For purposes of reducing heteroscedasticity, the assumption was made that the standard deviation of u_t was proportional to Y_t, so the whole equation was divided through by Y_t to get

$$\begin{aligned} \mathrm{APC}_t &= \frac{k_0}{Y_t} + k_1 r_t + k_2\left(\frac{W_t}{Y_t} - \frac{A_t}{Y_t}\right) + \sum_{j=0}^{q} k_{3+j} \frac{A_{t-j}}{Y_t} \\ &\quad + \sum_{j=0}^{n} w_j(z_{t-j} - q_{t-j}) + \sum_{j=0}^{n} \beta_j \left[(1 - \lambda)x_t^j + \sum_{i=1}^{m} (1 - D_t^i)s_{t-j}^i\right] + \epsilon_t, \end{aligned} \tag{B1}$$

where

$$z_{t-j} \equiv Y_{t-j}/Y_t \quad (\text{note: } z_t = 1 \text{ for all } t),$$

$$x_t^j = X_t^j/Y_t,$$

$$s_{t-j} = \frac{S_{t-j}}{Y_t},$$

$$q_{t-j} = s_{t-j} - \lambda x_t^j,$$

$$\epsilon_t = \frac{u_t}{Y_t}.$$

To estimate (B1), the assumption was made that both w_j and β_j follow third-degree polynomials in j:

$$\begin{aligned} w_j &= a_0 + a_1 j + a_2 j^2 + a_3 j^3, \\ \beta_j &= b_0 + b_1 j + b_2 j^2 + b_3 j^3. \end{aligned} \tag{B2}$$

The end-point constraints mentioned in the text ($w_{n+1} = \beta_{n+1} = 0$) are thus

$$\begin{aligned} a_0 + (n+1)a_1 + (n+1)^2 a_2 + (n+1)^3 a_3 &= 0, \\ b_0 + (n+1)b_1 + (n+1)^2 b_2 + (n+1)^3 b_3 &= 0. \end{aligned} \tag{B3}$$

Equations (B3) were used to eliminate the parameters a_0 and b_0. The adding-up constraint discussed (and rejected) in the text was thus

$$\begin{aligned} &\sum_{j=0}^{n} \{a_1[j - (n+1)] + a_2[j^2 - (n+1)^2] + a_3[j^3 - (n+1)^3]\}, \\ &= \sum_{j=0}^{n} \{b_1[j - (n+1)] + b_2[j^2 - (n+1)^2] + b_3[j^3 - (n+1)^3]\}, \end{aligned} \tag{B4}$$

which was used to eliminate the parameter a_1.

Finally, in estimating (B1), ϵ_t was assumed to follow a first-order autoregressive scheme, $\epsilon_t = \rho\epsilon_{t-1} + e_t$, where e_t is white noise. Estimation was by nonlinear least squares, which is equivalent to maximum likelihood if e_t is normally distributed. The function actually minimized was

$$\begin{aligned} \sum_{t=1}^{T} \Bigg\{ &\mathrm{APC}_t - \frac{k_0}{Y_t} - k_1 r_t - k_2\left(\frac{W_t - A_t}{Y_t}\right) \\ &- \sum_{j=0}^{q} k_{3+j} \frac{A_{t-j}}{Y_t} - \sum_{j=0}^{n} w_j(z_{t-j} - q_{t-j}) \\ &- \sum_{j=0}^{n} \beta_j \left[(1-\lambda)x_t^j - \sum_{i=1}^{m} (1 - D_t^i)s_{t-j}^i\right] - \rho\epsilon_{t-1} \Bigg\}^2, \end{aligned}$$

with all the above-mentioned definitions and parameter restrictions substituted in.

The estimating forms when regular income was further disaggregated were derived in precisely analogous ways from equations (16) and (18).

Appendix C

Calculation of the Real After-Tax Interest Rate

The real after-tax interest rate is defined as $r_t = i_t(1 - \tau_t) - \pi_t$, where i_t is the nominal interest rate, τ_t is the marginal tax rate, and π_t is the expected rate of inflation.

Nominal interest rate.—Four different nominal interest rates were tried: a corporate bond rate, the 4–6 month commercial paper rate, the 3-month Treasury bill rate, and a weighted average of rates received on various time and savings accounts. While all four led to very similar results, the Treasury bill rate gave the best fit and so was adopted.

Marginal tax rate.—The marginal tax rate was created from the average tax rate in the following way. Let $T(y)$ be the tax function facing an individual

taxpayer, and let $f(y)$ be the frequency distribution of income. The only directly observable tax rate is the aggregate average rate, which is $A = [\int T(y)f(y)dy]/[\int yf(y)dy]$.

Now suppose the whole income distribution shifts to the right with no change in its shape; that is, it shifts from $f(y)$ to $f[y(1-h)]$, where h connotes a "small" multiplicative shift. The average income is then $Y(h) = \int yf[y(1-h)]dy$, so that, for small shifts (h near zero), $dY/dh = -\int y^2 f'(y)dy$. Similarly, average tax payments after the shift are $T(h) = \int T(y)f[y(1-h)]dy$, so that $dT/dh = -\int yT(y)f'(y)dy$. The aggregate marginal tax rate, M, is the ratio of these: $M \equiv [\int yT(y)f'(y)dy]/[\int y^2 f'(y)dy]$.

To evaluate these integrals and obtain a closed expression for M/A, I adopted the following functional forms:

$$T(y) = ay^b \qquad (b > 1, a > 0),$$
$$f(y) = \gamma e^{-\gamma y} \qquad (\gamma > 0).$$

With these assumptions, the two ratios of integrals work out to be $A = (a\gamma^2/\gamma^{b+1})\Gamma(b+1)$, $M = (a\gamma^3/2\gamma^{b+2})\Gamma(b+2)$, where $\Gamma(n)$ is the "gamma function," namely, $\Gamma(n+1) = n\Gamma(n)$. Hence $M/A = (b+1)/2$, and using a tax elasticity of $b = 1.6$ gives $M = (1.3)A$.

The average tax rate was computed, quarter by quarter, by dividing the sum of federal and state-local personal tax and nontax payments by personal income excluding transfers (an approximation to the tax base). These are all official national income accounts series.

Expected rate of inflation.—Inflationary expectations were generated by a model based on what has been called "economically rational" expectations.[23] The idea is that agents, in informing themselves, begin with the data that are cheapest per unit of informational content and then proceed to process more costly data until the marginal cost and (expected) marginal benefits are equated. In this particular application, I assumed that consumers base their expectations of the inflation rate ($\dot{P}_t$) on its own past history and on the history of the growth rate of the money supply ($\dot{M}_t$). Thus I estimated an equation

$$\dot{P}_t = a_0 + \sum_{j=1}^{J} a_j \dot{P}_{t-j} + \sum_{i=0}^{I} b_i \dot{M}_{t-i} + e_t$$

on actual U.S. quarterly data, using the deflator for personal consumption expenditures for P_t and M_2 for M_t, and assumed that consumers used this equation to generate expectations. In estimation, I used the Almon (1965) lag technique with third-degree polynomials, no end-point constraints, and various choices for J and I. The best results were obtained with $J = 11$ and $I = 17$, namely[24]

$$\underset{(.38)}{\dot{P}_t = -.60} + \sum_{1}^{11} a_j \dot{P}_{t-j} + \sum_{0}^{17} b_i \dot{M}_{t-i}, \quad R^2 = .73, \qquad \text{D-W} = 1.98,$$

standard error = 1.42, mean of dependent variable = 3.35,

$$\underset{(.10)}{\Sigma a_j = .67} \qquad \underset{(.09)}{\Sigma b_i = .27}.$$

[23] See Feige and Pearce (1976).

[24] Standard errors are in parentheses. The D-W is the Durbin-Watson statistic. The period of estimation was 1951:3–1977:4, the longest period possible given the need for 17 lagged values of M.

References

Almon, Shirley. "The Distributed Lag between Capital Appropriations and Expenditures." *Econometrica* 33 (January 1965): 178–96.

Blinder, Alan S. "Distribution Effects and the Aggregate Consumption Function." *J.P.E.* 83, no. 3 (June 1975): 447–75.

———. "Intergenerational Transfers and Life Cycle Consumption." *A.E.R. Papers and Proc.* 66 (May 1976): 87–93.

Blinder, Alan S., and Solow, Robert M. "Analytical Foundations of Fiscal Policy." In *The Economics of Public Finance*, by Alan S. Blinder and others. Washington: Brookings Inst., 1974.

Boskin, Michael J. "Taxation, Saving, and the Rate of Interest." *J.P.E.* 86, no. 2, pt. 2 (April 1978): S3–S27.

Darby, Michael R. "The Allocation of Transitory Income among Consumers' Assets." *A.E.R.* 62 (December 1972): 928–41.

———. "Postwar U.S. Consumption, Consumer Expenditures, and Saving." *A.E.R. Papers and Proc.* 65 (May 1975): 217–22.

Dolde, Walter. "Forecasting the Consumption Effects of Stabilization Policies." *Internat. Econ. Rev.* 17 (June 1976): 431--46.

———. "Capital Markets and the Short Run Behavior of Life Cycle Savers." *J. Finance* 33 (May 1978): 413–28.

———. "Temporary Taxes as Macro-economic Stabilizers." *A.E.R. Papers and Proc.* 69 (May 1979): 81–85.

Eisner, Robert. "Fiscal and Monetary Policy Reconsidered." *A.E.R.* 59 (December 1969): 897–905.

Feige, Edgar L., and Pearce, Douglas K. "Economically Rational Expectations: Are Innovations in the Rate of Inflation Independent of Innovations in Measures of Monetary and Fiscal Policy?" *J.P.E.* 84, no. 3 (June 1976): 499–522.

Foley, Duncan K., and Hellwig, Martin F. "Asset Management with Trading Uncertainty." *Rev. Econ. Studies* 42 (July 1975): 327–46.

Friedman, Milton. *A Theory of the Consumption Function.* Princeton, N.J.: Princeton Univ. Press (for Nat. Bur. Econ. Res.), 1957.

Goldfeld, Stephen M., and Quandt, Richard E. *Nonlinear Methods in Econometrics.* Amsterdam: North-Holland, 1972.

Hall, Robert E. "Stochastic Implications of the Life Cycle–Permanent Income Hypothesis: Theory and Evidence." *J.P.E.* 86, no. 6 (December 1978): 971–87.

Houthakker, Hendrik S., and Taylor, Lester D. *Consumer Demand in the United States.* Cambridge, Mass.: Harvard Univ. Press, 1966.

Juster, F. Thomas. "A Note on Prospective 1977 Tax-Cuts and Consumer Spending." Unpublished paper, Univ. Michigan, January 1977.

Lucas, Robert E., Jr. "Econometric Policy Evaluation: A Critique." In *The Phillips Curve and Labor Markets*, edited by Karl Brunner and Allan H. Meltzer. Amsterdam: North-Holland, 1976.

Modigliani, Franco, and Brumberg, Richard E. "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data." In *Post-Keynesian Economics*, edited by Kenneth K. Kurihara. New Brunswick, N.J.: Rutgers Univ. Press, 1954.

Modigliani, Franco, and Steindel, Charles. "Is a Tax Rebate an Effective Tool for Stabilization Policy?" *Brookings Papers Econ. Activity*, no. 1 (1977), pp. 175–203.

Muth, John F. "Optimal Properties of Exponentially Weighted Forecasts." *J. American Statis. Assoc.* 55 (June 1960): 299–306.

Okun, Arthur M. "The Personal Tax Surcharge and Consumer Demand, 1968–70." *Brookings Papers Econ. Activity*, no. 1 (1971), pp. 167–211.

Sargent, Thomas J. "Rational Expectations, Econometric Exogeneity, and Consumption." *J.P.E.* 86, no. 4 (August 1978): 673–700.

Springer, William L. "Did the 1968 Surcharge Really Work?" *A.E.R.* 65 (September 1975): 644–59.

U.S. Bureau of Economic Analysis. *Survey of Current Business.* Washington: Government Printing Office, various issues.

Consumer Search with Uncertain Product Quality

John D. Hey

University of York

Chris J. McKenna

University College, Cardiff

This paper presents a unified model of consumer search when both prices and qualities are uncertain. It is assumed that price can be observed prior to purchase but that quality can be observed only after purchase and experience. This two-stage decision problem is of particular interest since, in practice, there is usually some connection between quality and price. Of the many interesting implications of this dependence explored in this paper is the result that the usual "reservation-price rule" may no longer be optimal; indeed, the optimal stopping set may take one of several alternative (and more realistic) forms.

I. Introduction

The economics of search has made considerable progress in recent years in describing and explaining the actions of economic agents operating in uncertain environments. Our concern in this paper is with the application of this large body of knowledge to the consumer-search problem. In its simplest form this problem examines the behavior of a consumer searching (at a constant cost per observation) over a known distribution of prices for the lowest (economically sensible) price for the purchase of one unit of some good. The

We would like to thank an anonymous referee and George J. Stigler for some very helpful comments on two earlier drafts of this paper. We are also grateful to Louis L. Wilde for useful comments on the original draft.

[*Journal of Political Economy*, 1981, vol. 89, no. 1]

solution, which is now well known, takes the form of a stopping rule uniquely determined by a "reservation price"; that is, the consumer behaves optimally by searching until a price lower than the reservation price is obtained.

In this simple form, and indeed in many of its more sophisticated generalizations, a crucial assumption is that search takes place over just one characteristic of the good—namely, its price. However, in practice, a good has several dimensions (of which price is but one); moreover, uncertainty surrounds not just the price dimension but these other dimensions as well. Of course, this observation would be trivial if information on these other aspects could be collected at the same time as information about price. However, in practice, some characteristics inevitably remain undiscovered until after purchase; thus, prior to purchase, residual uncertainty remains.

The purpose of this paper is to explore the consumer-search problem in this multidimensional form. In particular, we partition the set of characteristics into those which can be observed prior to purchase and those which can be observed only after purchase. To keep our exposition simple, we consider the simplest of all possible formulations: a two-dimensional case in which *price* is the dimension which can be observed prior to purchase and *quality* is the dimension observable only on purchase and experience.

Our extension makes the consumer's problem a two-stage one: The first stage relates to the price dimension—the consumer must decide what prices are "acceptable"; the second stage relates to the quality dimension—the consumer must decide what qualities are "acceptable." What makes the problem particularly interesting is that, in practice, these two dimensions—price and quality—tend to be related. This dependence means that the two stages of the consumer's problem are also related. Moreover, this dependence implies, as we shall show, that the optimal stopping rule at the first stage may well no longer be of the reservation-price form. For example, it may well be optimal for the consumer to search for expensive (rather than cheap) goods; alternatively, it could be optimal for the consumer to search for goods in some middle price range. Indeed, a variety of possible stopping rules may apply at the first stage; as we shall show, the form depends crucially on the form of the dependence between quality and price.

This two-stage problem belongs to a class of similar problems now beginning to appear in the literature. A similar consumer-search problem has been developed independently by Wilde (1977). In many respects his model is similar to ours, though it proceeds at a somewhat higher level of generality. Wilde's work follows from an earlier paper (Wilde 1979) on the parallel problem in the job-search area (a prob-

lem in which the wage is observed prior to accepting a job offer, and the nonpecuniary job characteristics are observed after experiencing the job). This latter problem was studied independently, and in a slightly different form, by McKenna (1978). Work has also been carried out in this area by Lippman and McCall (in press).

The paper is organized as follows: In the next section we introduce our notation and detail the assumptions of our model; the following section finds the solution to the consumer's two-stage problem and examines the properties of this solution; a concluding section reviews and appraises the key results.

II. The Model

We assume that the consumer-searcher is interested in two characteristics of the good—quality and price. We denote by Q the quality of the good as measured by the per period money value of the product's quality; we assume that the value of Q for a particular good can be observed only after purchase. The good yields Q for each period of its n-period lifetime; n is assumed to be the same for all goods. We denote by P the price of the good as measured by the price per unit per period of its lifetime (i.e., the per period "user cost" of the good).[1]

The consumer makes independent random selections from the joint distribution of P and Q; each drawing costs a (constant) amount c (> 0). The value of P is immediately revealed, while the value of Q is revealed only if purchase takes place. We assume that the searcher knows the joint distribution of P and Q over all goods: We denote by $g(\cdot)$ the marginal probability density function (pdf) of P, and by $f(\cdot \mid p)$ the conditional pdf of Q given p; we use $G(\cdot)$ and $F(\cdot \mid p)$ to denote the corresponding distribution functions and permit P and Q to take values on the positive real line.

The consumer faces an infinite horizon, discounts all values at the rate ρ ($0 < \rho < 1$), and wishes to have exactly one unit of the good in each period. Given these assumptions, the lifetime discounted value to the consumer is

$$\sum_{t=1}^{\infty} \rho^{t-1}(Q_t - P_t - S_t), \tag{1}$$

where Q_t and P_t are the quality and price of the good being consumed in period t, and S_t is the amount spent on search (possibly zero) in period t.

The individual's objective is to find, and follow, the strategy which

[1] Alternatively, the purchase price—payable at the time of purchase—is $\Sigma_{t=1}^{n}\rho^{t-1}P$ (where ρ, see below, is the discount rate).

maximizes the expected value of the expression (1). The model is built around the following assumed sequence of events:

1. The consumer searches, at a cost c per search, over values of P until an acceptable value of P is obtained. We assume that this search process is instantaneous and that no passage of time occurs.[2] Thus the consumer can obtain instantaneously n "quotes" at a total cost nc, whether the searching is done sequentially or not.
2. The consumer then purchases one unit of the good.
3. After purchase, the consumer experiences the good and thereby discovers the value of Q associated with the good. (We assume that all uncertainties are resolved after purchase.)
4. Having discovered Q, the consumer decides whether to keep the good for the whole of its n-period lifetime or to dispose of the good (at zero cost) at some point within its lifetime. If the latter option is used, the individual returns to stage 1.
5. If the good is kept until the end of its lifetime, the consumer must decide whether Q was sufficiently good for it to be repurchased. If so, the good is continually repurchased thereafter; otherwise, search continues again at stage 1.

The model is appropriate for regularly purchased nondurable goods for which the amount consumed per period is roughly constant, such as bread, cereals, eggs, washing powder, newspapers, and toilet paper. However, the purchase of some durable items such as televisions, cars, and washing machines may also be described in our framework.[3]

III. The Solution

The consumer's problem is a two-stage one: The first stage relates to the price dimension, the second stage to the quality dimension. Our solution similarly proceeds in two stages.

We begin with the first stage (that described by 1 in the sequence of events above), namely, the stage relating to the price dimension. The consumer's problem at this stage is to decide which price quotes are

[2] This is crucial in view of our earlier assumption that the consumer requires exactly one unit of the good per period. In contrast, Wilde (1977) postulates a time-consuming search process and assumes that the consumer derives no utility in periods when he does not have a unit of the good and derives utility $U(P, Q)$ in other periods.

[3] The repurchase of durable goods, however, may be made after a much longer period of consumption. For our model to be valid in describing purchase decisions of this type requires that we assume unchanging market conditions over long periods.

acceptable (in that purchase should take place at such quotes if obtained) and which price quotes are unacceptable (in that search should continue). In other words, the consumer must decide on the optimal acceptance set for prices. To find this optimal set, we first begin with an arbitrary acceptance set which we denote by $\mathscr{P}$ ($\subset \mathscr{R}^+$). This set $\mathscr{P}$ defines a stopping rule applicable at stage 1 as follows:

$$\begin{aligned} &\text{if } P \in \mathscr{P}, \text{ stop searching and purchase good at } P; \\ &\text{if } P \notin \mathscr{P}, \text{ continue searching.} \end{aligned} \tag{2}$$

Here P denotes the latest price quote obtained. We wish to find the best such set, which we denote by $\mathscr{P}^*$, in the sense that it maximizes the expected value of the objective function (1). To find this $\mathscr{P}^*$ we proceed as follows. Consider the consumer starting from a position in which no acceptable price has yet been obtained; suppose that the consumer uses the arbitrary acceptance set $\mathscr{P}$ with respect to the first purchase decision but thereafter behaves optimally with respect to all future decisions (including those relating to any future return to stage 1). Let $V(\mathscr{P})$ denote the expected value of the objective function following this sequence of events. Clearly $\mathscr{P}^*$, by definition, is a set for which $V(\mathscr{P})$ attains its maximum value; if we denote this latter by V^*, we have

$$\begin{aligned} &V^* \geq V(\mathscr{P}) \qquad \text{for all } \mathscr{P} \in 2^{\mathscr{R}^+}, \\ &V^* = V(\mathscr{P}^*). \end{aligned} \tag{3}$$

We will demonstrate that $\mathscr{P}^*$ need not be determined by a unique reservation price (i.e., $\mathscr{P}^*$ need not be of the form $\{P; P \leq P^*\}$). This is because the direct attraction of low prices may be more than offset by the indirect detraction of associated low qualities. In mathematical terminology, the return function from accepting P need not be monotonically decreasing in P because of the relationship between Q and P in the joint distribution. We pursue this formally below.

We now turn to the second stage (that described by 4 and 5 in the sequence of events above), namely, the stage relating to the quality dimension. The consumer's problem at this stage is to decide what qualities are acceptable, both in terms of keeping the good until the end of its lifetime and in terms of repurchasing that good at the end of its lifetime. In contrast with the decision at stage 1, these acceptance decisions do have the reservation property—the higher the Q, the better off the individual.[4]

[4] In contrast with the first stage where a higher P may make the individual better or worse off. For a formal demonstration of the textual assertion see Hey (1979) and/or Theorem 1 of Wilde (1979).

We wish to determine the optimal reservation qualities; to find these we first begin with arbitrary reservation levels defined as follows. Let q_i denote an arbitrary reservation quality used at the end of period i following the purchase of the good ($i = 1, \ldots, n - 1$).[5] Using these arbitrary values, the decision at stage 4 is as follows:

$$\begin{aligned} &\text{if } Q < q_i, \text{ dispose of the good at the end of period } i \text{ (if not already disposed of) and return to stage 1;} \\ &\text{if } Q \geq q_i, \text{ keep the good (at least until the end of period } i + 1). \end{aligned} \tag{4}$$

Here Q denotes the quality of the good purchased.

Further, let q_n denote an arbitrary reservation quality used at the end of period n (at the end of the life of the good). Using this arbitrary value, the decision at stage 5 is as follows:

$$\begin{aligned} &\text{if } Q < q_n, \text{ do not repurchase the good; instead return to stage 1;} \\ &\text{if } Q \geq q_n, \text{ continue repurchasing that good forever.} \end{aligned} \tag{5}$$

Although not essential, particularly for arbitrary q_i, it would seem appropriate to assume that $q_i \leq q_{i+1}$ ($i = 1, \ldots, n - 1$).

To find the consumer's optimal strategy, we need to find the best set of q_i ($i = 1, \ldots, n$), which we denote by q_i^* ($i = 1, \ldots, n$), in the sense that this set maximizes the expected value of the objective function (1). To find these q_i^* we proceed as follows. Consider the consumer starting from a position in which a purchase at price P has just been made. Suppose that the consumer uses arbitrary reservation qualities q_i ($i = 1, \ldots, n$) with respect to this purchase but thereafter behaves optimally with respect to all future decisions (including those relating to any future returns to stages 4 and 5). Let $W(P; q_1, \ldots, q_n)$ denote the expected value of the objective function following this sequence of events. Clearly, q_i^* ($i = 1, \ldots, n$) is a set for which $W(P; q_1, \ldots, q_n)$ attains its maximum value (for given P); if we denote this latter by $W^*(P)$, we have

$$\begin{aligned} &W^*(P) \geq W(P; q_1, \ldots, q_n) \quad \text{for all } q_i \in \mathscr{R}^+ \ (i = 1, \ldots, n), \\ &W^*(P) = W(P; q_1^*, \ldots, q_n^*). \end{aligned} \tag{6}$$

This completes our notational preliminaries, and we can now proceed to an investigation of $\mathscr{P}^*$ and the q_i^*.

If the stopping set $\mathscr{P}$ is used at stage 1, we have

$$V(\mathscr{P}) = -c + V(\mathscr{P})[1 - \int_{\mathscr{P}} g(p)dp] + \int_{\mathscr{P}} W^*(p)g(p)dp. \tag{7}$$

[5] We exclude $i = 0$ on grounds of realism; its inclusion would not materially affect any of the results.

This equation states that, starting from a position in which no acceptable price quote has been found, the individual searches once more at cost c and either receives an unacceptable quote (returning to "square one") or receives an acceptable quote with expected value of $W^*(p)$.[6] Equation (7) gives:

$$V(\mathcal{P}) = [-c + \textstyle\int_{\mathcal{P}} W^*(p)g(p)dp]/\int_{\mathcal{P}} g(p)dp. \tag{8}$$

Now, it can easily be shown (Hey 1979) that an optimal set is given by

$$\mathcal{P}^* = \{P;\ W^*(P) \geq V^*\}, \text{ where } c = \textstyle\int_{\mathcal{P}^*} [W^*(p) - V^*]g(p)dp. \tag{9}$$

The first of these simply states that a price quote should be accepted if the expected value (of the objective function) obtainable by accepting it is at least as large as that obtainable by rejecting it. The value of V^* is determined by (9); we assume that this is positive (otherwise it is not optimal for the consumer to initiate search for the good).

Examination of (9) shows that the form of the set $\mathcal{P}^*$ depends crucially on the behavior of the function W^*. The familiar case is when W^* is a monotonically decreasing function (i.e., when higher prices unambiguously make the consumer worse off); in this case, as can be seen from (9), the set $\mathcal{P}^*$ is given by $\{P; P \leq P^*\}$, where P^* is the familiar reservation price. However, in our model there is nothing to guarantee that W^* is monotonically decreasing; indeed, W^* may take a variety of different forms, depending on the relationship between quality and price in the joint distribution. In the extreme, it is possible (although unlikely) that quality increases so fast with price that the function W^* is monotonically increasing; in such a case the acceptance set $\mathcal{P}^*$ would take the form $\{P; P \geq P^*\}$ where P^* is a lower reservation price—a buy-expensive policy would be optimal. The general case (one in which W^* has both increasing and decreasing segments) is illustrated in figure 1.

As the form of the function W^* is crucial to the consumer's optimal behavior, we now proceed to an investigation of its properties. From our story of consumer behavior, we have

$$\begin{aligned} W(p; q_1, \ldots, q_n) = {} & \left[-p(1-\rho^n)/(1-\rho) + \int_0^\infty qf(q\,|\,p)dq\right] \qquad (10) \\ & + \Sigma_{i=1}^{n-1}\ \rho^i\left[V^*\int_{q_{i-1}}^{q_i} f(q\,|\,p)dq + \int_{q_i}^\infty qf(q\,|\,p)dq\right] \\ & + \rho^n V^*\int_{q_{n-1}}^{q_n} f(q\,|\,p)dq + \rho^n\int_{q_n}^\infty [(q-p)/(1-\rho)]f(q\,|\,p)dq \end{aligned}$$

(where $q_0 \equiv 0$).

[6] Note that the search process is assumed to be instantaneous and thus the expected values following the search are not discounted; see also n. 2 above. Note further that if $\mathcal{P} = \phi$, then (7) yields $V(\phi) = -c + V(\phi)$, which (if $c > 0$) implies that $V(\phi)$ is minus infinity.

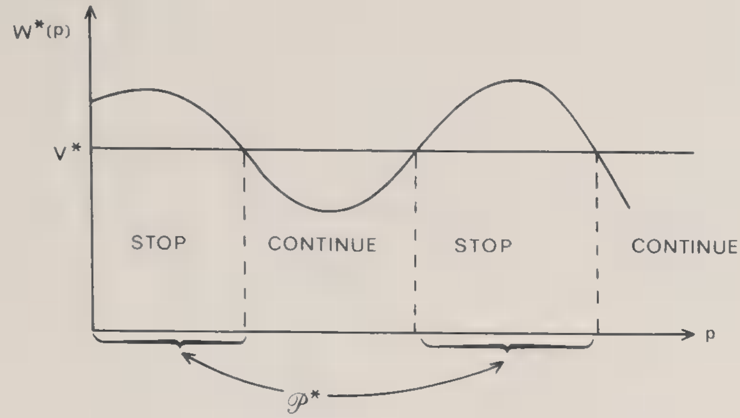


FIG. 1.—Derivation of the set $\mathscr{P}^*$ in the general case

The optimal reservation qualities at the end of periods 1 to $(n - 1)$ —the q_i^* $(i = 1, \ldots, n - 1)$—are found by setting $\partial W/\partial q_i = 0$. It is easy to show that this implies that

$$q_i^* = V^*(1 - \rho) \qquad (i = 1, \ldots, n - 1). \tag{11}$$

Two things of interest flow from this: First, $q_1^* = \ldots = q_{n-1}^*$, which means that if the good's quality is sufficiently high for it not to be thrown away at the end of the first period of its lifetime, it should be kept for the remainder of its lifetime. (This result is a consequence of our assumed constancy of the per period quality rate; if, in contrast, quality were to decline as the good got older, this strong result would no longer hold.) Second, $q_1^* = V^*(1 - \rho)$, which means that a good should be thrown away (before the end of its useful life) if and only if its quality is lower than the per period value obtainable from starting search afresh.

The optimal reservation quality at the end of the good's lifetime is found by setting $\partial W/\partial q_n = 0$; this implies that

$$q_n^* = V^*(1 - \rho) + p, \tag{12}$$

which states that the consumer should *re*purchase the good if and only if its quality net of price is at least as large as the expected value obtainable from starting search afresh. Both this, and the earlier result (11), accord with intuition.

If we now substitute (11) and (12) into (10), we get an expression for W^*; after some simplification, this yields

$$\begin{aligned} W^*(p) = \rho V^* - [p(1 - \rho^n)/(1 - \rho)] + \int_0^\infty [1 - F(q \mid p)]dq \\ + [\rho(1 - \rho^{n-1})/(1 - \rho)] \int_{q_1^*}^\infty [1 - F(q \mid p)]dq \\ + [\rho^n/(1 - \rho)] \int_{q_n^*}^\infty [1 - F(q \mid p)]dq \end{aligned} \tag{13}$$

(where, in view of the equality of the q_i^* for $i = 1, \ldots, n - 1$, we simply work with q_1^* and q_n^*).

Equations (9), (11), (12), and (13) solve for V^*, $W^*(\cdot)$, q_1^*, q_n^*, and $\mathcal{P}^*$ in terms of $c, \rho, n, f(\cdot \mid \cdot)$, and $g(\cdot)$; they provide the complete solution to our model.

The form of (13) can now be investigated; from it and (12) we have

$$
\begin{aligned}
W^{*\prime}(p) = {} & -(1 - \rho^n)/(1 - \rho) - \int_0^\infty \frac{\partial F(q \mid p)}{\partial p} dq \\
& - [\rho(1 - \rho^{n-1})/(1 - \rho)] \int_{q_1^*}^\infty \frac{\partial F(q \mid p)}{\partial p} dq \\
& - [\rho^n/(1 - \rho)] \int_{q_n^*}^\infty \frac{\partial F(q \mid p)}{\partial p} dq \\
& - [\rho^n/(1 - \rho)][1 - F(q_n^* \mid p)].
\end{aligned}
\tag{14}
$$

Inspection of this shows that $W^{*\prime}$ is unambiguously negative if $\partial F(q \mid p)/\partial p$ is positive everywhere; however, this latter condition implies that $F(\cdot \mid p)$ is stochastically decreasing in p—or, in economic terms, that quality is stochastically decreasing in price. This is highly unlikely in practice. A more realistic case is to assume that quality generally increases with price so that $F(\cdot \mid p)$ is stochastically increasing in p (i.e., $F(t \mid p_1) \geq F(t \mid p_2)$ for all t whenever $p_1 < p_2$; see Hadar and Russell [1969]). However, this restriction is not sufficient to sign $W^{*\prime}$. On reflection, this is not a surprising result; one can imagine cases where quality increases very strongly with price (so that $W^{*\prime}$ is positive) and, also, cases where quality increases very slowly with price (so that $W^{*\prime}$ is negative—the direct price effect outweighing the indirect quality effect). In general, it is possible that $W^{*\prime}$ has both positive and negative sections, in which case the stopping rule at stage 1 would be of a form similar to that shown in figure 1.

Wilde (1977, 1979) also discusses similar possibilities (see, in particular, Wilde [1979], n. 11). Moreover, because his model proceeds at a rather higher level of generality than our own (namely, his objective function is not simply linear in quality and price), he finds restrictions (sufficient to sign his equivalent of our $W^{*\prime}$) on the joint distribution virtually impossible to obtain. Instead, he simply assumes that the distribution is such that the return function (W^*) is monotonically decreasing.

However, some further insight can be gained as follows.[7] Denote by Q_p the (random) quality associated with the price p and denote the

[7] Although our model assumes that "utility" is linear in price and quality, our results would still go through if utility was a separable function of price and quality if the distributions $f(\cdot \mid \cdot)$ and $g(\cdot)$ were reinterpreted appropriately.

distribution function of Q_p by $F_p(\cdot) \equiv F(\cdot \mid p)$. Suppose that the conditional distributions are such that

$$F_p(q) = F_0(q - kp), \tag{15}$$

that is, Q_p is a translation of Q_0. (Thus the distribution of Q_p is the same as that of Q_0 shifted to the right by a distance kp.) It is a simple matter to show that

$$\partial F(q \mid p)/\partial p = -kf(q \mid p). \tag{16}$$

If this is substituted into (14), we find, after some simplification, that

$$\begin{aligned} W^{*\prime}(p) = (k - 1)[1 - \rho^n F(q_n^* \mid p)]/(1 - \rho) \\ - k\rho(1 - \rho^{n-1})F(q_1^* \mid p)/(1 - \rho). \end{aligned} \tag{17}$$

There is an important special case in which the second term in (17) disappears; this is when $F(q_1^* \mid p) \equiv 0$, or, in economic terms, when there is no possibility of finding a quality so bad that it pays to throw the good away before the end of its lifetime.[8] In such a case, equation (17) implies that

$$W^{*\prime}(p) \gtreqless 0 \text{ according as } k \gtreqless 1. \tag{18}$$

This result makes good sense: It states that the return function is an increasing (decreasing) function of price if $q - p$ (the quality net of price) is an increasing (decreasing) function of price. Moreover, if $q - p$ is independent of price (that is, if $k = 1$), W^* is horizontal (at the level V^*), and the consumer simply accepts the first price quote received whatever its value.

The general case is when $F(q_1^* \mid p) > 0$, that is, when there exists a positive probability of getting a good whose quality is so low that it pays to throw it away and start search afresh. In this general case, the simple result (18) no longer holds. However, as inspection of (17) shows, if k is less than or equal to one, then $W^{*\prime}$ is strictly negative. Thus, if "net quality" $(q - p)$ decreases with price, the return function is monotonically decreasing and a buy-cheap policy remains optimal. However, k larger than 1 is not now sufficient for $W^{*\prime}$ to be positive, because the (nonzero) possibility of getting a "lemon" destroys the generally greater attractiveness of higher-price goods.[9] However, by differentiating (17) (and using [16]) it can be shown that k greater than 1 is a sufficient condition for $W^{*\prime\prime}$ to be positive. Furthermore, from (17) it can be seen that $\lim_{p\to\infty} W^{*\prime}(p) = (k - 1)/(1 - \rho) > 0$ if $k >$

[8] This effectively is the case considered by Wilde (1977).

[9] A sufficient condition for $W^{*\prime}$ to be positive is that k is at least as large as $(1 - \rho^n)/(1 - \rho)$, so that net quality increases sufficiently fast with price to outweigh the nonzero lemon possibility.

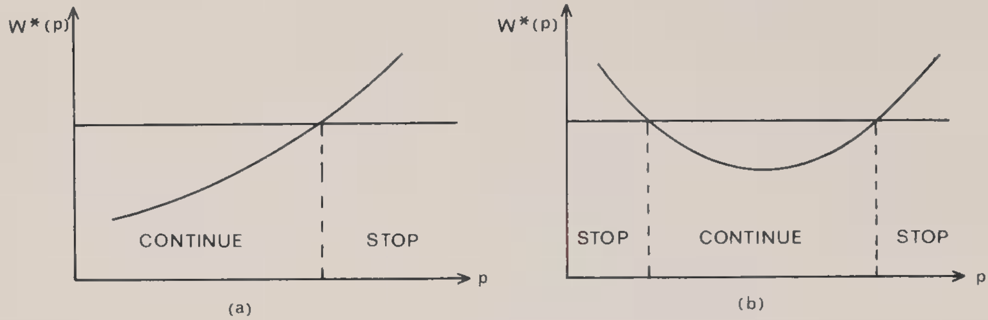


Fig. 2.—Derivation of $\mathscr{P}^*$ when (15) holds with $k > 1$

1.[10] These results imply that if k is greater than 1 (i.e., if net quality increases with price), then $W^{*\prime}$ is monotonically increasing and that even if it is negative at lower values of p, it must become positive at some higher value. Thus, if k is greater than 1, one of the two possibilities pictured in figure 2 must exist.

Case b in figure 2 is particularly interesting, implying that the consumer should search until either a very low, or a very high, price quote is obtained. Goods at intermediate prices should not be purchased. This result follows from the combination of net quality increasing with price and the lemon possibility: High prices are attractive because the associated high qualities are a sufficiently great incentive to outweigh the possible lemon effect; low prices are attractive despite the associated low qualities because of the small loss involved in throwing away a lemon which cost little in the first place; intermediate prices are not attractive because they fall between these two stools.

The results discussed in the paragraphs above are, of course, specific to the conditional distribution formulation given by (15). However, these results do yield some insight into the general considerations underlying the form of the function W^*; clearly important is the (stochastic) relationship between net quality $(q - p)$ and price. Alternative formulations can easily be envisaged. For example, imagine that the conditional distribution of net quality increases at an increasing rate with price up to some intermediate price and thereafter increases at a decreasing rate. In such a case, W^* may be an increasing function for low p and a decreasing function for high p; the price-acceptance set would then consist of intermediate prices, both very low and very high prices being rejected in favor of further search.

Clearly, a whole variety of possibilities exist: Depending on the

[10] Since $F_p(q)$, as defined in (15), approaches zero as p approaches infinity for any finite q.

conditional distribution of quality given price, buy cheap, buy expensive, buy medium, or various combinations of these are all plausible buying strategies. Thus "gaps" in the stopping set $\mathcal{P}^*$ can be quite consistent with optimal behavior.

Before concluding we note the following comparative static results. They can be derived from equations (9), (11), (12), and (13); the detailed proofs can be found in Hey and McKenna (1979).

1. Each of the following makes the consumer-searcher better off in that V^* and $W^*(p)$ increase: (i) a decrease in the search cost c, (ii) a decrease in the lifetime of the good n, (iii) a rightward shift of the conditional distributions of Q_p, (iv) a rightward shift of the marginal distributions of P if W^* is a monotonically increasing function, (v) a leftward shift of the marginal distribution of P if W^* is a monotonically decreasing function, (vi) an increase in risk (in the Rothschild and Stiglitz [1970] sense) in the conditional distributions of Q_p. Consequently the consumer becomes more choosy in that q_1^* and q_n^* increase and the set $\mathcal{P}^*$ gets smaller.

2. The expressions $V^*(1 - \rho)$ and $W^*(p)(1 - \rho)$ increase if ρ increases, that is, if the consumer places greater weight on the future; consequently, the consumer gets more choosy in the present in that q_1^* and q_n^* increase and the set $\mathcal{P}^*$ gets smaller.

All of these results accord with intuition, with the possible exception of that relating to n. However, when it is remembered that P is the price per period of the good's lifetime, it does make sense to conclude that a smaller n is better, in that it reduces the time spent "locked in" to an unsatisfactory good.

IV. Summary and Conclusions

This paper has presented a unified model of consumer search with both price and quality uncertainty. Because price is revealed prior to purchase while quality is not revealed until after purchase, the consumer's problem is essentially a two-stage problem. The first stage relates to the price and involves choosing an acceptance set for prices. The second stage relates to the quality, and involves deciding whether a particular quality is sufficiently good (*a*) to retain it after purchase and (*b*) to continue buying it thereafter. The problem is made interesting and nontrivial by the dependence of quality on price.

The major result is the finding that the price-acceptance set is not necessarily connected and that the conventional reservation price rule is optimal only in some circumstances. In the general case, depending on the stochastic relationship between quality and price, buy cheap, buy expensive, buy medium, or various combinations of these are all plausible buying strategies.

In the form presented in the paper, the consumer's utility is assumed to be linear in price and quality. However, all our results still hold if utility is a separable function of price and quality, as long as the distributions of price and quality are reinterpreted appropriately (as distributions of the utility of price and the utility of quality). Of course, in such a generalization optimal behavior would vary from consumer to consumer, depending on the differences between their utility functions. It would be of interest to explore the implications for price/quality equilibria in markets resulting from such behavior. It is to be hoped that this paper has laid a firm foundation for such an exploration.

References

Hadar, Josef, and Russell, William R. "Rules for Ordering Uncertain Prospects." *A.E.R.* 59 (March 1969): 25–34.

Hey, John D. "A Simple Generalised Stopping Rule." *Econ. Letters* 2 (July 1979): 115–20.

Hey, John D., and McKenna, Chris J. "Consumer Search with Uncertain Product Quality." Discussion paper no. 33, Univ. York, Inst. Soc. and Econ. Res./Dept. Econ., 1979.

Lippman, Steven A., and McCall, John J. "The Economics of Belated Information." *Internat. Econ. Rev.* (in press).

McKenna, Chris J. "A Theory of Contractual Decision under Uncertainty." Discussion paper no. 27, Univ. York, Inst. Soc. and Econ. Res./Dept. Econ., 1978.

Rothschild, Michael, and Stiglitz, Joseph E. "Increasing Risk. I: A Definition." *J. Econ. Theory* 2 (September 1970): 225–43.

Wilde, Louis L. "On the Formal Theory of Inspection and Evaluation in Product Markets." Mimeographed. Pasadena: California Inst. Tech., 1977.

———. "An Information-theoretic Approach to Job Quits." In *Studies in the Economics of Search*, edited by Steven A. Lippman and John J. McCall. Amsterdam: North-Holland, 1979.

Commodity-Choice Behavior with Pigeons as Subjects

Raymond C. Battalio and John H. Kagel

Texas A&M University

Howard Rachlin

State University of New York at Stony Brook

Leonard Green

Washington University

Starting from an initial (baseline) budget line, income-compensated price changes always resulted in substitution effects consistent with the Slutsky-Hicks theory. This behavior cannot be explained by a simple random-behavior model. Similar changes in relative prices that did not originate from the initial (baseline) budget line resulted in "undersubstitution effects": The composition of consumption changed in the expected direction, but the magnitude of change was not large enough to be consistent with the initial commodity bundle chosen. These undersubstitution effects are not explainable by shifting preference patterns or anchoring effects found in inconsistent choice sequences with human subjects.

This paper reports tests of the negativity of the Slutsky-substitution effect in commodity-choice experiments with pigeons as subjects. Starting from a given (baseline) budget line, income-compensated

Research support from the National Science Foundation is gratefully acknowledged. An earlier version of this paper was presented at the 1979 Summer Econometric Society meetings in Montreal. We thank the editor and the referee for helpful comments. The usual caveat applies.

[*Journal of Political Economy*, 1981, vol. 89, no. 1]

increases or decreases in relative commodity prices always resulted in substitution effects consistent with the Slutsky-Hicks theory. These results, which replicate earlier studies using rats as subjects (Kagel et al. 1975, 1980), are not explainable by simple random-behavior models proposed in the literature (Becker 1962; Chant 1963).

Similar changes in relative prices that did not originate from the baseline resulted in "undersubstitution effects"; that is, the composition of consumption tended to change in the expected direction, but the changes were not large enough to be consistent with the original (baseline) commodity bundle consumed. These undersubstitution effects are not explainable in terms of shifting preference patterns or anchoring effects (Tversky and Kahneman 1974) found in inconsistent choice sequences with humans as subjects. Further experiments will be required to identify the basis for this behavior.

Data on learning patterns as subjects adjust to changes in experimental conditions are also reported. Questions of interest here concern identifying uniformities in both the speed of the adjustment process and its pattern, whether it can be characterized by concepts of habit formation or stock-adjustment processes (Houthakker and Taylor 1970). The nature of the learning process is, of course, strictly an empirical matter in terms of the static Slutsky-Hicks theory.

Space considerations do not permit a detailed discussion of the rationale behind using nonhuman subjects to investigate economic theories (see Kagel and Battalio [1980] for this argument). For the more skeptical reader we note that if one defines economics as ". . . the study of the allocation of scarce resources among unlimited and competing uses" (Rees 1968), then animal psychologists, ecologists, and biologists have been involved in studying economic behavior for some time now (Rapport and Turner 1977; Hirshleifer 1978; Rachlin 1980). It is but a small step to take the technologies of these related disciplines and apply them to behavior of interest to economists, for example, commodity-choice behavior. At a minimum, such studies expand considerably the scope for comparative economic analysis. At a maximum they provide a laboratory for identifying, testing, and better understanding general laws of economic behavior. Use of this laboratory is predicated on the fact that behavior as well as structure vary continuously across species, and that principles of economic behavior would be unique among behavioral principles if they did not apply, with some variation, of course, to the behavior of nonhumans.

The plan of the paper is as follows: Section I describes the hypotheses tested and the procedures used to operationalize the budget constraint. Sections II and III report the two sets of experiments conducted: Both comparative static and temporal patterns of behavior are examined. We end with a discussion which summarizes the test outcomes and their implications for future research and returns to

the question of the relevance of the experiments for understanding human economic activity.

I. Hypotheses Tested and Experimental Procedures Employed

A necessary condition that all members of the class of Slutsky-Hicks demand functions must satisfy is that for income-compensated price changes (of the Slutsky type) the consumer purchase a commodity bundle that was previously unattainable. More precisely,

$$\text{if } p^i \neq p^j \text{ and } p^j x^i = p^j x^j \text{ then } (p^j - p^i)(x^j - x^i) < 0, \tag{1}$$

where p^i and x^i represent vectors of prices and quantities of goods purchased in period i. Condition (1), also referred to as the weak axiom of revealed preference, provides the basis for a direct consistency test of the theory (Hicks 1956).

Condition (1) was tested by having food- and water-deprived pigeons choose between the consumption of these two commodities. The subjects were placed in an experimental chamber for a fixed period of time each day. Each chamber contained a food hopper and water reservoir located on the front panel. Access to the food and water was controlled by separate variable time schedules. Under a variable time schedule commodities become available for a fixed period of time every α seconds on the average, where α is determined as part of the experimental design. Only one of the variable time schedules was in operation at a given time, and subjects could switch between the food and water schedule by a single peck on a control key. Pecks on this key were not required to continue food or water deliveries, but only to change which commodity would be delivered. Color-coded lights were used to signal whether the food or water schedule was in effect. The price of each commodity was varied by altering the average time between deliveries of that commodity (access times remaining constant), while the income constraint was the total time available for the delivery of the two goods.

Under these procedures time spent under the food and water schedules (T_F, T_W) produces commodities at an average rate of

$$C_F = \frac{1}{\alpha_F} T_F, \qquad C_W = \frac{1}{\alpha_W} T_W, \tag{2}$$

where α_F, α_W are the average values of the variable time schedules, and C_F and C_W are food and water consumption measured in number of payoffs. Since time required to switch between schedules is sufficiently small that it can be ignored for all practical purposes, substituting (2) into the total time (T_T) constraint

$$T_T = T_W + T_F \tag{3}$$

gives an expected value for the budget constraint of

$$\alpha_F C_F + \alpha_W C_W = T_T. \tag{4}$$

From (4) it is clear that increasing (decreasing) the average time interval between deliveries of a commodity has the same impact as a price increase (decrease) in the usual specification of the budget constraint. Analogously, increasing (decreasing) the total time available for delivery of the goods increases (decreases) nominal income (see Becker [1971] for the comparative statics of this problem).

To establish conditions for testing the negativity of the Slutsky-substitution effect, consumption patterns were first recorded under a set of initial (baseline) price (α_F^i, α_W^i) and income (T_T^i) conditions. The price of one of the commodities was then increased, accompanied by a corresponding decrease in the price of the other commodity so as to achieve, as nearly as possible, perfect compensation

$$\alpha_F^j C_F^i + \alpha_W^j C_W^i = \alpha_F^j C_F^j + \alpha_W^j C_W^j, \tag{5}$$

where the superscripts refer to the initial (i) and price-change period (j). Since commodity consumption varied somewhat from day to day under constant experimental conditions, the median bundle consumed during the last 5 days of a condition served as the reference point for compensation. Total time available for consumption remained constant across experimental conditions, as did the amounts of food and water delivered per payoff.

Consistent choice behavior (a negative Slutsky-substitution effect) requires that

$$T_T^i - (\alpha_F^i C_F^j + \alpha_W^i C_W^j) < 0. \tag{6}$$

That is, with the compensated change in relative prices the consumer purchases a commodity bundle in the price-change period that it could not afford to purchase during baseline. Since, as already noted, consumption was not the same from day to day under constant experimental conditions, t-tests applied to the mean value of (6) calculated over the last 10 days of the price-change period were used to determine consistency.

In cases where (6) is negative, a simple random-behavior process could explain the results (Becker 1962; Chant 1963). For example, if the pigeons randomly choose the amount of time spent receiving each of the commodities, and total time remains constant across experimental conditions, then reducing the time between food deliveries, which corresponds to a reduction of food costs, will result in increased food consumption. Correspondingly, increasing the time between water payoffs, which increases its price, would result in a decrease in average water consumption. Under random behavior the

expected value of time spent receiving each of the commodities would be

$$\frac{\alpha_k C_k}{T_T} = q_k \qquad \begin{matrix} k = F, W \\ 0 \leq q_k \leq 1 \end{matrix}, \tag{7}$$

where q_k is constant across experimental conditions.[1]

Chant has characterized consumers who behave in this fashion as "money deciders." In the case where $q_k = .5$, the pigeons would be spending an equal amount of time collecting each of the commodities. In Chant's terminology they would be "impulsive money deciders." For the more general case, where the pigeons spend a constant but unequal proportion of time collecting each of the commodities, they would be characterized as "proportionately inert money deciders." In cases where (6) is negative, we will test whether behavior is significantly different from (7) using t-tests for paired comparisons and appropriate analysis-of-variance procedures for comparing across several conditions. Rejection of (7) suggests that something other than a simple random behavioral process is guiding choices.

II. Experiment 1

Procedures

Four male White Carneaux pigeons with no previous experimental history served as subjects (S's). Subjects were completely dependent on food and water consumed during experimental sessions with no access to these or any other commodities in their home cages. Experimental sessions were run once a day 7 days a week starting at approximately the same time each day. Consequently, subjects' body weight varied somewhat between conditions depending on consumption patterns.

All S's had extensive training with the procedures prior to the start of the experiment. Experimental conditions were changed when inspection of graphs of the data indicated an absence of any significant drift in consumption patterns, although at times conditions were maintained for considerably longer periods simply to see what would happen.

Food payoffs consisted of 8 seconds access to a food hopper containing mixed pigeon grains. Fluid payoffs consisted of 3 seconds access to a reservoir containing tap water. Access to the food hopper

[1] Negative values for (6) rule out the cases of "inert" or "impulsive goods deciders" (Becker 1962; Chant 1963). Of course, behavior in accordance with (7) is observationally indistinguishable from behavior being motivated by a Cobb-Douglas utility function (Becker 1962).

and water reservoir was blocked at all other times. Time spent eating and drinking did not count against total time available for delivery of commodities (T_T), which was 60 minutes for all S's.[2]

For three of the four S's, food and water were delivered on the average at 60-second intervals during the initial (baseline) period. This was followed by an increase in the relative price of food. Baseline conditions were then reinstated, followed by a decrease in relative food prices, which was followed by a return to baseline conditions. (Table 1 shows the sequence and magnitude of relative price changes.) To account for any shifts in preference structure, the median commodity bundle taken in the last 5 days of each baseline served as the reference point for the Slutsky-compensated price change of the following period. A fourth subject, S 85, started with food payoffs coming twice as often as water deliveries (relative price of food equal to .5), followed by a Slutsky-compensated increase in the price of food. Baseline conditions faced by the other birds were implemented in period 3 for this subject. Since the mean commodity bundle purchased during period 3 could very nearly have been bought during the previous period, we compare these two points also.

Comparative Static Results

The effects of the Slutsky-compensated price changes are reported in table 1. The results of the increase in food price relative to baseline conditions are reported between conditions 1 and 2 for S's 14, 46, and 50 and between conditions 2 and 3 for S 85. In all cases the weak-axiom test statistic (6) has the correct sign and is statistically significant at the .05 level for three of the four S's. Price elasticities for water ranged from a low of $-.07$ for S 50 to a high of $-.81$ for S 14, averaging $-.38$ across all four birds. These price elasticities are a bit larger than the highest values reported for food-fluid experiments with rats (Kagel et al. 1980). The increase in the relative price of food for S 85 between periods 1 and 2 also resulted in substitution effects in the expected direction.

Decreases in the relative price of food for S's 14, 46, and 50 in period 4 resulted in a negative (correct) sign for the weak-axiom test statistic, which was significant at the .05 level or better in all cases. Price elasticities here ranged from a low of $-.10$ for S 50 to a high of $-.24$ for S 14, averaging $-.16$ across S's.

Replication of baseline conditions following the price-change

[2] Initial values were determined so that S's could maintain healthy body weights from food and water consumed within the sessions. Values were determined on the basis of past experiments and experience during the training periods. Additional procedural details may be found in the Appendix.

TABLE 1

RESULTS OF CONSISTENCY TESTS: EXPERIMENT 1

SUBJECT AND EXPERIMENTAL CONDITION	α_F/α_W	WEAK-AXIOM TEST (t-Statistic)[a]	MEAN PAYOFFS[b] (S_M) Food	Water	F-TEST FOR EQUALITY OF BASELINES (Marginal Significance Level)
Subject 14:					
1	1.0	−636 (2.44)**	46.4 (1.54)	13.6 (1.54)	
2	2.88		39.5 (2.20)	31.1 (6.55)	
3	1.0	−312 (3.26)***	37.1 (3.05)	23.3 (2.92)	3.36** (.05)
4	.28		48.7 (2.21)	16.5 (.82)	
5	1.0		42.4 (1.31)	18.2 (1.16)	
Subject 46:					
1	1.0	−498 (2.74)**	34.2 (1.47)	25.8 (1.47)	
2	2.86		32.8 (1.63)	35.5 (4.65)	
3	1.0	−288 (2.79)**	37.2 (1.58)	22.9 (1.57)	.94 (.40)
4	.22		43.7 (2.08)	21.1 (.41)	
5	1.0		36.0 (1.61)	23.4 (1.64)	
Subject 50:					
1	1.0	−48 (.30)	41.1 (1.62)	19.1 (1.55)	
2	4.0		40.0 (.98)	20.8 (3.66)	
3	1.0	−168 (2.43)**	42.6 (.82)	17.7 (.92)	2.39 (.11)
4	.32		47.2 (1.58)	15.6 (.50)	
5	1.0		44.7 (1.04)	16.0 (.94)	
Subject 85:					
1	.5	−551 (1.87)*	33.5 (1.59)	33.4 (.82)	
2	4.5	−798 (3.88)***	31.3 (1.33)	42.0 (4.74)	N.A.
3	1.0		31.4 (3.28)	28.2 (3.31)	

NOTE.—α_F/α_W = relative food cost; N.A. = not applicable; S_M = standard error of the mean.
[a] The t-tests are two tailed with 9 degrees of freedom.
[b] Calculated over the last 10 days of each period.
* Significant at the 10 percent level.
** Significant at the 5 percent level.
*** Significant at the 1 percent level.

TABLE 2

Test of Simple Random-Behavior Model

Subject and Grouping*	Income Spent on Food (%)	Experimental Condition	Relative Price of Food	F-Test for Equal Percentage Hypothesis (Marginal Significance Level)
Subject 14:				
A	.79	2	2.88	
A	.77	1	1.00	
A, B	.71	5	1.00	13.4† (.001)
B	.62	3	1.00	
C	.46	4	.28	
Subject 46:				
A	.73	2	2.86	
B	.62	3	1.00	
B	.60	5	1.00	32.7 (.001)
B	.57	1	1.00	
C	.32	4	.22	
Subject 50:				
A	.89	2	4.00	
B	.75	5	1.00	
B	.71	3	1.00	55.4 (.001)
B	.69	1	1.00	
C	.50	4	.32	
Subject 85:				
A	.78	2	4.50	
B	.52	3	1.00	34.8 (.001)
C	.34	1	.50	

* Conditions with the same grouping letter have the same percent of total expenditure devoted to food according to Duncan's multiple-range test (α = .05).

† F-statistic excluding period 1 equals 14.7.

periods showed consumption patterns to be reversible, at least directionally, in all cases. Preferences were sufficiently stable that for S's 46 and 50 there were no significant differences in consumption patterns between baseline periods. While consumption for S 14 in period 1 is found to differ significantly from that of periods 3 and 5 (at the .05 level using Duncan's multiple-range test [Winer 1971]), there are no differences between periods 3 and 5. We note that the ratio of food to water consumption for S 14 in period 1 was the highest value found across all S's in both experiments under baseline prices.

Table 2 reports tests of the simple random-behavior model (7). In all cases F-tests show that we can decisively reject the hypothesis that the percentage of total income spent on commodities was the same across differing experimental conditions. Using the letters A, B, and C to indicate experimental conditions that can be grouped together as having the same percentage of income spent on food (at the .05 significance level according to a Duncan multiple-range test), we see that the price-change periods are responsible for the differences in

each case. For S's 46 and 50 the percent of total income spent on food remained constant across baseline periods (conditions 1, 3, and 5) but increased significantly in period 2 when relative food prices increased and dropped significantly in period 4 when food prices dropped. Subjects 14 and 85 show similar response patterns with the exception, already noted, that consumption patterns differed between baseline periods for S 14. In all cases demand is far too inelastic for consistency to be attributed to the fact that S's spent a constant proportion of their time collecting each of the commodities. These pigeons cannot be characterized as either "impulsive" or "proportionately inert" money deciders.

Adjustment Patterns

The comparative static analysis tells us nothing about the learning path as subjects adjust to changes in relative prices. To investigate this we fit, within a given experimental condition, the following hyperbolic function to the commodity whose relative price decreased:

$$C_{kt} = a + b(1/t),$$

$$C_{kt} = \text{consumption of commodity } k \text{ on day } t\text{, commodity } k \text{ being the good whose relative price decreased,} \tag{8}$$

$$t = \text{time in days within a given experimental condition.}$$

The function (8) builds in the assumption that consumption approaches a steady-state value as time passes and there are no further changes in experimental conditions. We fit these functions because (1) in terms of conducting the experiments, conditions were only changed when inspection of the data suggested no significant drift in consumption patterns, and (2) graphs of the data suggested that in most cases we did not need to use more complicated functions to adequately represent the time path.

The coefficient b of (8) is readily interpretable in terms of a number of concepts within the economics and psychology literature. If b is not significantly different from zero we have, in effect, instantaneous, or at least very rapid, adjustment to the price changes, relative to the day-to-day variability in consumption patterns. Fitting (8) to the commodity whose relative price decreased under a given experimental condition and finding a negative value for b would imply that subjects responded gradually to the change in relative prices, increasing consumption of the now cheaper good with the passage of time. Such behavior would be consistent with concepts of habit formation found useful in describing human consumption for a number of nondurable goods (Houthakker and Taylor 1970; Phlips 1974).

The presence of habit formation would also be consistent with psychological notions that it takes S's time to perceive the changes in relative prices and even more time to learn the implications of responding to these changes.

In contrast, a positive sign for b in (8) would indicate that the immediate response to a decrease in relative prices was to consume more of the commodity than with the passage of time. This sort of response pattern goes under the heading of a stock-adjustment effect within the Houthakker-Taylor framework. It might be present in our data for any number of reasons.

In general there is no a priori basis for deciding whether habit-formation or stock-adjustment effects will predominate for a particular commodity. For example, Houthakker and Taylor (1970) found that housing exhibited strong habit-formation effects, while domestic services and water use showed stock-adjustment effects generally considered to be characteristic of consumer durable goods. The empirical work in psychology shows much the same diversity since learning curves of almost every conceivable shape have been reported in the literature (Mazur and Hastie 1978).

Figure 1 shows the daily water consumption for condition 2 when the relative price of water decreased. The mean consumption of water during the last 10 days of the previous period is also shown as an open circle on the water axis. The estimated b values and the Durbin-Watson statistics obtained from the least-squares estimates of equation (8) are also given for each subject.[3] The data show little trend and a great deal of variability. This is evident in the least-squares estimates of the time-trend coefficients, b, which are not statistically significant for any of the S's. Note, however, that in each case b is positive in sign, so that what little trend there is in the data is inconsistent with the notion that habit formation dominates choices. That is, there is a tendency to have more substitution into water during the first few days than is shown in the comparative static comparisons.

Figure 2 shows the daily food consumption for condition 4 when the relative price of food decreased. Once again there appears to be little trend, a result captured in the t-test ratios for the least-squares estimates of b, which are insignificant in all cases, although the signs of the coefficients are somewhat more favorable to the habit-formation hypothesis.

The absence of significant time-trend coefficients, in conjunction with the data reported in table 1, implies that S's adjusted very quickly to the changes in relative prices, at least within the time frame for

[3] The breaks in the sequential data plotted in figs. 1–4 correspond to days when the equipment malfunctioned. Dummy variables are used in these cases when adjusting (8) to the data.

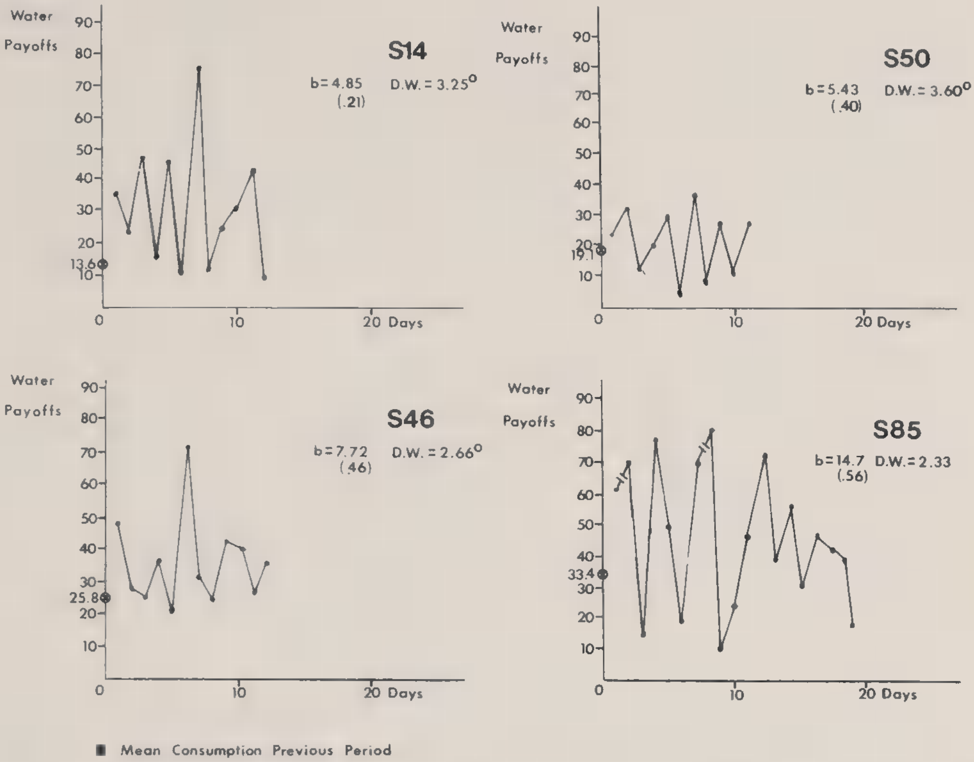


FIG. 1.—Temporal pattern of water consumption following a decrease in its relative price: experiment 1, condition 2. Broken lines indicate days with equipment failures. Estimated values of the time trend coefficient, *b*, with the corresponding *t*-statistic reported in brackets are given for each subject along with the value of the Durbin-Watson statistic. The symbols +, *, and ** indicate statistical significance at the 10, 5, and 1 percent levels, respectively. The symbols – and o indicate the Durbin-Watson test was indeterminate or there was an insufficient number of observations for the test.

which we have data (daily totals) and relative to the day-to-day variability in the data. For example, in going from condition 1 to 2, *S* 14 more than doubled water intake but has a time-trend coefficient with a *t*-statistic considerably less than 1.0. The frequent trips to the cafeteria windows (*S*'s received 60 or more payoffs per session), the low search costs involved in choosing within the confines of the experimental test chamber, and the physiological limits on substitutability imposed by the commodities employed all facilitate *S*'s adjusting rapidly to the changes in experimental conditions. The data here indicate that at least some of the subjects took advantage of these conditions.

The Durbin-Watson statistics, while indicating a tendency to negative first-order autocorrelation in the residuals, do not show a serious problem in these respects. Such a problem might be anticipated from the physiological interdependencies between food and water, which would result in the pigeon's getting caught short on food or water for a given day, only to be compensated for in the next day's consumption

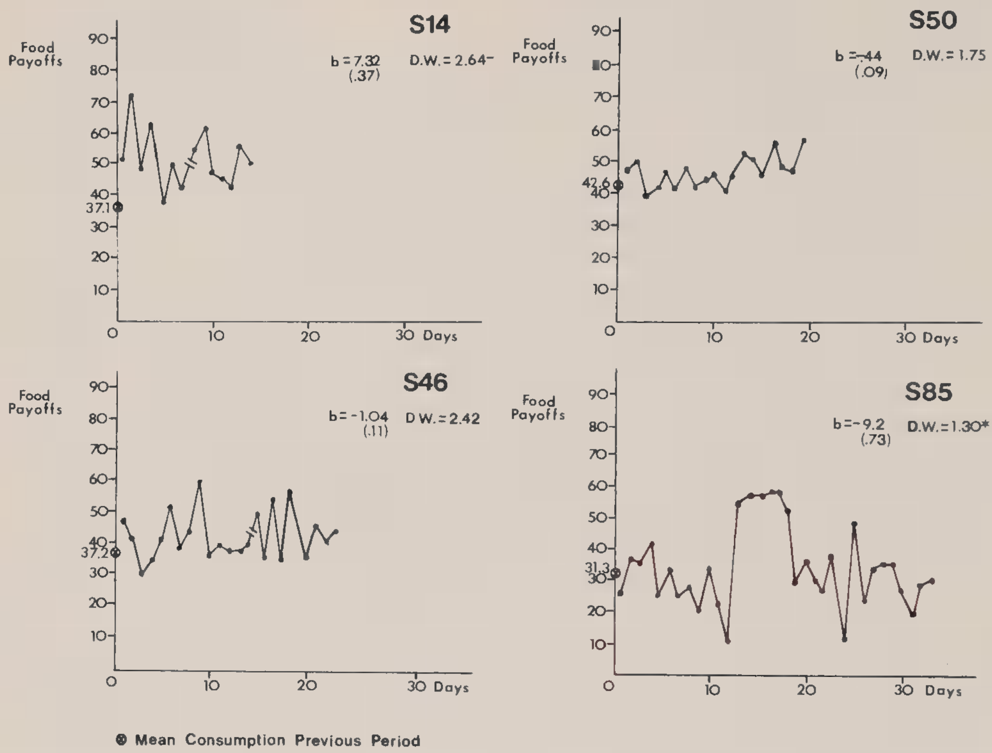


FIG. 2.—Temporal pattern of food consumption following a decrease in its relative price: experiment 1, condition 4. Notation is the same as in figure 1.

bundle.[4] The data indicate, however, that this did not happen. Subjects either had sufficiently good "planning," given the frequent trips to the cafeteria window in a given day, or sufficiently frequent cross-overs between food and water schedules so as not to generate imbalances strong enough to dominate the residuals. Durbin-Watson statistics computed on the basis of fitting (8) to the data from the baseline periods show similar results.

III. Results for Experiment 2

Procedures

Essentially the same procedures were employed as in experiment 1. The major difference was the sequence of price changes.

Four experimentally naive male White Carneaux pigeons served as subjects. Initial (baseline) conditions were the same as for *S*'s 14, 46, and 50 in experiment 1 with the exception that *S* 38 obtained 13

[4] We are assuming that *S*'s can and will correct imbalances within one day. This assumption is based on data for days following equipment failures (e.g., a malfunctioning food dispenser) which show that, for rats and pigeons alike, excluding data for the day of the failure and the following day generally leaves the time trend unaffected.

seconds access with each food payoff and 6 seconds access to the water reservoir. The difference in price-change sequence from experiment 1 consisted of (1) decreasing relative food prices first and (2) not returning directly to baseline conditions but following this with a compensated increase in food prices relative to baseline before returning to baseline conditions. (Table 3 shows the sequence and magnitude of price changes.)

Comparative Static Results

Table 3 reports the effects of the compensated price changes. The decrease in relative food prices in period 2 led to statistically significant substitution effects in the expected direction in all cases. The price elasticities of demand for food ranged from a low of $-.21$ for *S* 38 to a high of $-.51$ for *S* 36, averaging $-.32$ across *S*'s. Applying t-tests to the percentage of total time spent obtaining food in periods 1 and 2, we again reject the hypothesis that the simple random-behavior model (7) explains any of the data: Demand is far too inelastic in this case also.

The compensated increase in relative food prices in period 3 does show some surprises, however. In all cases the relative price of food was greater than during the initial baseline period. However, the predicted magnitude of change in consumption did not materialize for any of the *S*'s. Although the composition of consumption (ratio of food to water payoffs) changed in the "right" direction from period 2 to 3 for three of the four *S*'s, in no case was the change sufficiently large to result in a correct sign for the weak-axiom test statistic (6). In contrast, comparable increases in relative food prices starting from baseline conditions in experiment 1 resulted in a correct sign for the test statistic in all cases. Further, a comparable increase in relative food prices for *S* 85 in period 2 of experiment 1, which started from a relative price ratio similar to that faced by *S*'s 36–39 in period 2 of this experiment, did result in consistent substitution effects significant at the 10 percent level.[5]

One potential explanation for the behavior observed in period 3 is that preferences had shifted. The return to baseline conditions in period 4 enables us to test this hypothesis. For the one subject, *S* 36, whose behavior is clearly inconsistent in period 3, preferences have clearly shifted ($t = 5.15$, df $= 18$, $P < .001$).[6] For the other three *S*'s showing nonexistent or undersubstitution effects, the hypothesis receives only weak support. Although food consumption has in-

[5] Real income was somewhat higher for *S* 85 in period 1.
[6] $P < .001$ = probability that we can reject the null hypothesis of no effect.

TABLE 3

RESULTS OF CONSISTENCY TESTS: EXPERIMENT 2

Subject and Experimental Condition	α_F/α_W	Weak-Axiom Test (t-Statistic)[a]	Mean Payoffs[b] (S_M): Food	Mean Payoffs[b] (S_M): Water	t-Test for Equality of Baselines (Marginal Significance Level)[c]
Subject 36:					
1	1.0	−1032 (11.3)***	24.1 (1.26)	35.9 (1.26)	
2	.22	546 (3.25)***	47.5 (1.89)	29.7 (.37)	5.15*** (.001)
3	9.02		27.6 (.37)	23.3 (3.15)	
4	1.0		33.2 (1.24)	26.8 (1.24)	
Subject 37:					
1	1.0	−312 (4.24)***	41.2 (1.47)	18.8 (1.47)	
2	.50	318 (1.88)*	49.6 (2.46)	15.6 (1.27)	.93 (.37)
3	4.0		42.1 (.90)	12.6 (3.71)	
4	1.0		43.8 (2.39)	16.2 (2.39)	
Subject 38:					
1	1.0	−228 (3.90)***	45.0 (1.30)	15.1 (1.35)	
2	.53	54 (1.03)	51.2 (2.05)	12.6 (1.09)	.78 (.45)
3	2.60		45.1 (.67)	14.0 (1.48)	
4	1.0		47.0 (2.21)	13.0 (2.21)	
Subject 39:					
1	1.0	−144 (2.51)**	39.9 (2.00)	20.2 (2.01)	
2	.53	0.0 (0.0)	47.0 (1.58)	15.4 (1.06)	1.78* (.09)
3	2.40		43.1 (.74)	16.9 (1.78)	
4	1.0		44.1 (1.25)	15.9 (1.25)	

NOTE.—α_F/α_W = relative food cost; S_M = standard error of the mean.
[a] The t-tests are two tailed with 9 degrees of freedom.
[b] Calculated over the last 10 days of each period.
[c] The t-tests are two tailed with 18 degrees of freedom.
* Significant at the 10 percent level.
** Significant at the 5 percent level.
*** Significant at the 1 percent level.

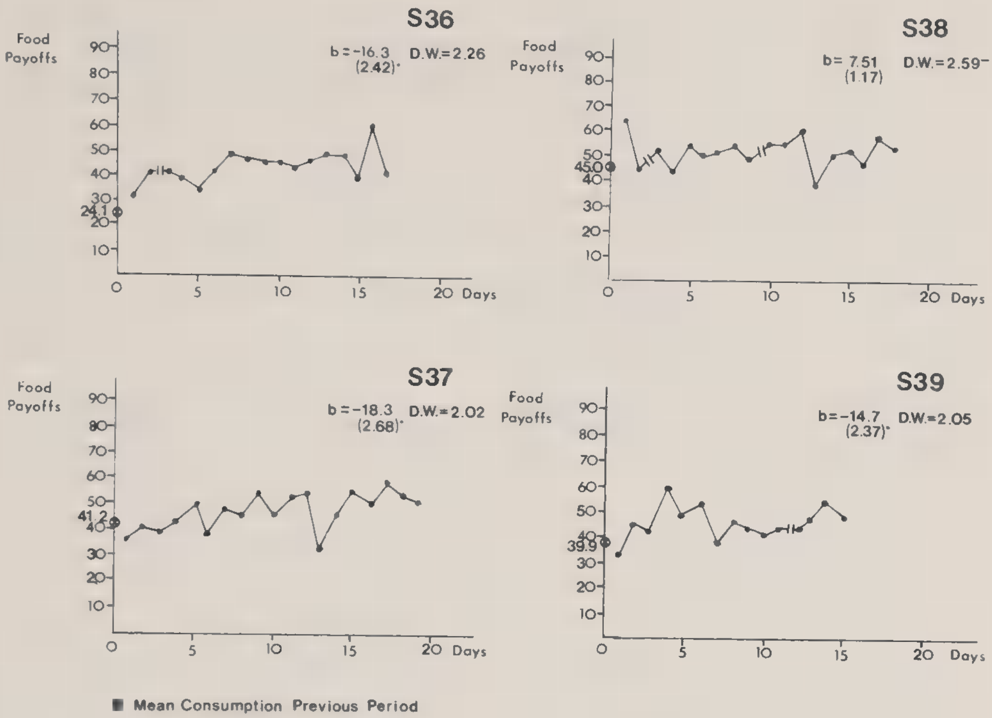


FIG. 3.—Temporal pattern of food consumption following a decrease in its relative price: experiment 2, condition 2. Notation is the same as in figure 1.

creased in period 4 relative to the initial period for all three, in no case does the increase seem to be sufficient, by itself, to explain the undersubstitution effects. The magnitude of the change in relative prices, and the fact that we are dealing with Slutsky-compensated price changes which, if anything, overestimate the Hicksian substitution effect (Friedman 1976), would seem to more than compensate for any small shifts in preference structure that might have occurred.[7] Finally, the behavior reported in period 3 is inconsistent with the random-response model (7) also.

Adjustment Patterns

Figure 3 shows food consumption in period 2 following the decrease in its relative price. The data show more of a positive time trend, consistent with the notion of habit formation, than comparable observations from experiment 1. Estimated values of the time-trend coefficient, b, from equation (8) confirm this as we have negative coefficients in three of the four cases, each one of which is significant at the 5 percent level. These negative coefficients indicate that subjects responded slowly to the decrease in relative food prices, consistent with notions of habit formation.

[7] We are assuming here that both food and water are normal goods.

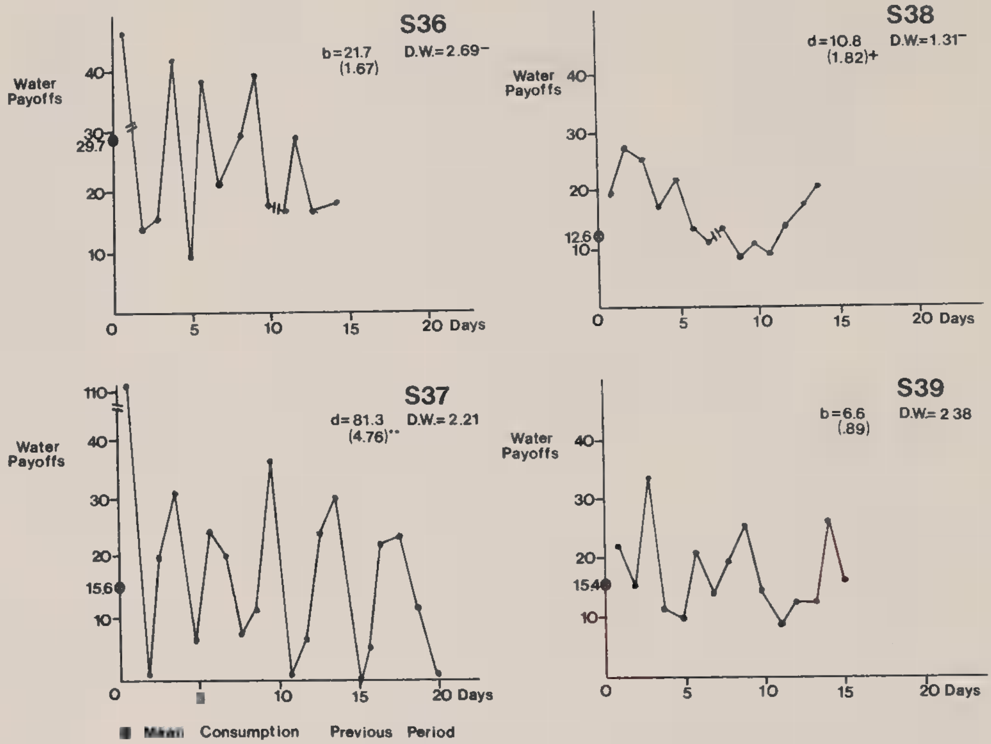


FIG. 4.—Temporal pattern of water consumption following a decrease in its relative price: experiment 2, condition 3. Notation is the same as in figure 1.

In contrast, the data for period 3, shown in figure 4, during which the relative price of water decreased, show stock-adjustment effects in all cases, being statistically significant at the 1 percent level for *S* 37 and at the 10 percent level for *S* 38. Thus, the initial response to the decrease in the price of water was greater at the beginning of this period than at the end. In fact, for all *S*'s the commodity bundle purchased on the first day of period 3 satisfied the basic consistency test (1). Further, for *S*'s 37–39, who did not shift preferences between baseline periods, the weak-axiom test statistic calculated over the first 3 days of period 3 has the correct sign, although it is somewhat smaller in magnitude than values observed under comparable conditions (period 2) in experiment 1. These data would appear to rule out arguments that the undersubstitution effects found in period 3 were a result of not maintaining experimental conditions for a sufficiently long period of time and/or that *S*'s failed to correctly process the magnitude of the changes in relative prices.

The adjustment patterns in the two experiments are qualitatively similar. Pooling the results from the two experiments shows an asymmetric response to increases and decreases in relative food prices. For decreases in prices we find a tendency to habit formation, or increased food consumption with the passage of time. For increases

in relative food prices the data consistently show anti–habit formation, more water (less food) consumed immediately following the price change than later on. These conclusions are reached whether we combine the t-test outcomes for the price-change periods ($Z = -2.28$, $P < .03$ for the t-values associated with b in periods when relative food prices decreased; $Z = 8.0, P < .001$ for the t-values associated with b in periods when relative water prices decreased) or use a χ^2 analysis based on the signs of the time-trend coefficients ($\chi^2 = 9.6$, df $= 1, P < .01$).[8]

The asymmetric response to increases and decreases in relative food prices is quite unexpected. One possible explanation lies in physiological asymmetries with respect to how animals process food and water. Food can be stored in the form of enhanced body weight, while the experimental procedures do not permit maintaining comparable stocks of water. Thus, in the case of increased food prices, there is a stock of food available to buffer large and immediate increases in water consumption. But with decreases in the price of food, there is not a stock of water to produce similar results. Further, food-buffer stocks were larger in experiment 2 at the time food prices increased since it came immediately after a large decrease in food prices in the previous period. This might have contributed to the stronger stock-adjustment effects found in this experiment.

Although adjustment patterns in the two experiments are qualitatively similar, a question arises as to what factors are responsible for the quantitative differences in the t-ratios and the size of the time-trend coefficients between the two experiments. The size of the reductions in relative food prices during condition 2 of experiment 2 is comparable to that faced by subjects in experiment 1 (see tables 1 and 3). Further, the length of the baseline periods preceding these price changes was similar in the two experiments. This would seem to rule out arguments that these quantitative differences result from the

[8] Under the hypothesis that the mean value for the t-statistic in the population is zero, the statistic,

$$z = \frac{\Sigma t_j}{\Sigma[f_j/(f_j - 2)]},$$

where f_j represents the degrees of freedom associated with t_j, has a sampling distribution which is approximately $N(0, 1)$ (Winer 1971). The picture gets a bit fuzzier if we include periods of return to baseline (conditions 3 and 5 in experiment 1, condition 4 in experiment 2) to the analysis. The Z-statistic for t-values when relative food prices decreased loses significance ($Z = -1.12, P = .26$), while for periods when relative water prices decreased the significance level drops ($Z = 2.60, P < .01$). Also, the significance level for the χ^2 test drops ($\chi^2 = 4.5$, df $= 1, P < .04$). The primary reason for this is that the sign of the b coefficient associated with food consumption in period 4 of experiment 1 is positive for S's 37–39, while to be more consistent with the price change periods it should be negative since food prices decreased going into this period.

magnitude of price changes or the length of time prices had remained stable. This leaves individual subject differences or differences in procedures (sequence effects or some undetected change) between the two experiments as the most likely factors responsible for the differences.

Attributing part of the differences to variability in individual subject responses receives some support. The learning-curve literature in psychology shows a lot of variability in response patterns using single-subject data, with some individual learning curves being very irregular or varying greatly from the general form for a given task (Mazur and Hastie 1978). Examination of the magnitude of the estimates of b within each experiment reveals a similar variability in behavior. For example, in both experiments one of the four subjects has a positive point estimate when the relative price of food decreased. Similar differences in individual subject responses are, of course, not detectable in aggregate per capita data such as Houthakker and Taylor employed. However, the consistent negative signs of the time-trend coefficients for decreases in relative food prices and positive signs for increases in relative food prices (decreases in the relative price of water) suggest that an aggregate per capita analysis, under conditions where all subjects faced the same relative price changes, would clearly exhibit these patterns.

Finally we note that the values for the Durbin-Watson statistics reported in figures 3 and 4 are comparable to those found in experiment 1. While there is a tendency to negative first-order autocorrelation in the residuals, these effects are rarely strong. Even though S's would readily consume more at the end of a session if given the opportunity, they managed to keep food and fluid intake relatively well balanced on a daily basis.

IV. Summary and Discussion

Using figure 5 we can readily summarize the consistency test results. Starting from some initial (baseline) budget line (line 1), with consumption bundle A, Slutsky-compensated increases or decreases in relative prices result in consistent and statistically significant substitution effects (points B and C, respectively). This result replicates similar tests of the Slutsky-Hicks theory using rats pressing levers and choosing between food and water, or flavored fluids with ad lib. access to food and water (Kagel et al. 1975, 1980). These results suggest the universality and robustness of one of the fundamental "truths" of the static theory of consumer demand, that compensated price increases will bring about a decrease in quantity demanded.

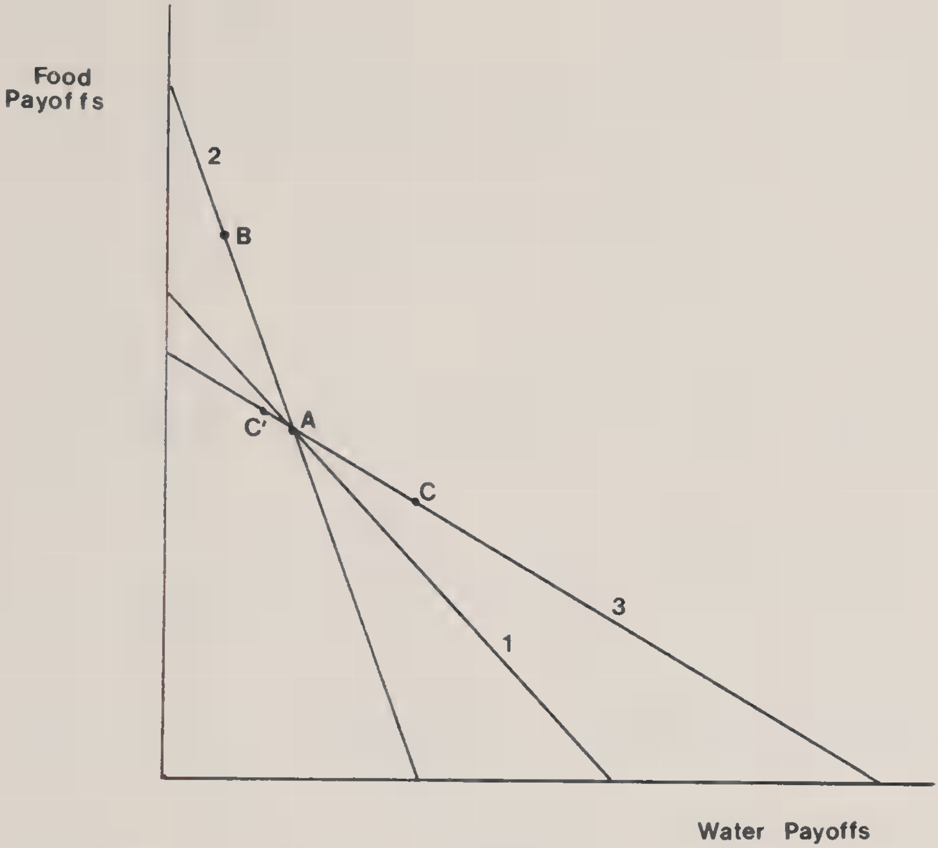


FIG. 5.—Summary of consistency-test results in terms of commodity-choice space. Numbered lines are budget lines under differing relative prices with capital letters characterizing choices. See text for details.

Further, this truth does not rest on simple random-behavior models suggested by Becker (1962) and Chant (1963) but apparently reflects a basic behavioral principle.

Large rotations in the budget line which involved swinging from line 2 to 3 without stopping off at baseline resulted in inconsistent choice behavior (point *C'*) relative to the initial commodity bundle consumed. While *S*'s tended to alter their composition of consumption in the expected direction, they did not change enough to be consistent with the initial commodity bundle consumed.[9] Efforts to explain these undersubstitution effects in terms of shifting preference structures and/or delayed reactions (learning) in response to the change in relative prices did not prove successful.

Although it would be nice to have an explanation for the inconsis-

[9] We searched our rat data for similar rotations in the budget line. In the one case this happened the results are similar to those reported here. This was overlooked at the time since further rotation of the budget line in the same direction resulted in additional changes in consumption so that the weak-axiom test statistic was satisfied.

tencies reported here in terms of established patterns of behavior derived from the choice experiment literature, we have not found one which is completely consistent. For example, anchoring effects, the fact that for humans estimates of quantities (Tversky and Kahneman 1974) and strengths of sensory phenomena (Stevens 1975) are greatly affected by the starting point used in making these estimates, might be used to explain the differences in going from budget line 2 to 3 directly rather than from 1 to 3. But anchoring, unless conjoined with a hypothesis specifying some minimum difference between points before it takes effect, would also seem to imply systematic biases in consumption patterns in returning to baseline conditions following a relative price change, something we did not observe in experiment 1. Explaining the inconsistencies in terms of changes in the animal's body weight resulting from the relative price changes runs into similar problems, although we cannot rule out more complex physiological interaction effects. Perhaps the behavior simply results from *S*'s inexperience, so that once having experienced a compensated price change starting from baseline prices, the same commodity bundle would be chosen starting from elsewhere in the choice space. Identification of the factors underlying the inconsistencies must wait further experimentation. Identifying these factors would also help clarify the behavioral mechanism underlying consistent choices.

Looking at day-to-day adjustments in consumption patterns resulting from changes in relative prices we find a number of cases in which adjustments were completed within the first observation period and other cases with statistically significant learning effects. Individual subject differences rather than the size of the price changes or other environmental factors were cited as one of the factors behind these differences. The fact that the adjustment coefficients tended to have the same sign whether they were statistically significant or not suggests that an aggregate analysis under conditions where all subjects faced the same price changes would produce statistically significant learning effects. Unanticipated asymmetries in the adjustment path resulting from increases and decreases in relative food prices were also found. We are looking at adjustment patterns from the rat studies in an effort to confirm these regularities.

Questions naturally arise concerning the relevance of the animal behavior reported here to understanding analogous human behavior, the behavior of primary interest to most economists. Objections might be raised on two grounds. One is that there are essential differences between the choice conditions (environmental conditions) the pigeons face and conditions facing human consumers operating in national economic systems. The pigeons are not given money (or

tokens, as in some human economic experiments) and told what prices are, but are simply faced with sequences of situations from which they must deduce what the rules of the game are. One might argue that this results in the pigeons facing a more difficult situation than is assumed in ordinary consumer-demand theory. Second, one might object that there are essential differences in the choice process itself, on what goes on inside the pigeon's brain, so to speak, relative to human consumers. The sequential way in which subjects learn about changes in relative prices and the apparent sequential nature of choices even under steady-state conditions stand in marked contrast to the textbook characterization of a consumer who maximizes a utility function subject to a budget constraint at a point in time, apparently deciding on all purchases at once. We deal with these objections below.

There is no question that the choice conditions the pigeons faced here have no exact physical analogue to choice conditions outside the laboratory for humans, or for pigeons in the wild for that matter. What an experiment aims for is to develop procedures in the laboratory which provide a logically consistent interpretation for the concepts and relationships of the theory under investigation. Within the context of consumer-demand theory, this involves developing procedures that alter the choice set in the same way that the symbols commonly designated "price" and "income" alter choices in the general theory. This is true of any application of the theory, be it experimental or nonexperimental, using human or nonhuman subjects, so that there are always some differences in environmental conditions in different applications of the theory, and these differences may affect decision outcomes.

If the theory is to have any generality it must hold under a variety of such choice conditions. Consequently, confirmations and disconfirmations are relevant to evaluating the truth status of the general theory under any and all logically consistent interpretations.[10] Further, it is often the case that the most preferred conditions for testing a theory will be "atypical." For example, one of the crucial experiments of relativity theory is to determine if stars behind the edge of the sun appear to be displaced from their known positions during a total solar eclipse. One does not object to such experimental observations simply on the grounds that they do not correspond to "normal" conditions (Smith 1980). Similarly, perfect compensation,

[10] Identifying differences in behavior under different interpretations of the general theory sets off a search for the factors responsible for these differences and eventually results in modifications of the general theory to account for these factors. Unfortunately, this is often an exceedingly slow process.

which is required to directly apply the weak-axiom test (1), rarely occurs in nature but has an important role to play in better understanding commodity-choice behavior. Establishing procedures for implementing these conditions using nonhumans is considerably less costly than with humans, while introducing tokens or some other conditioned reinforcer to serve as a medium of exchange and unit accounting, although technically feasible, would only complicate the experiment with little likelihood of seriously affecting behavior. In fact, we might argue that having subjects learn the rules of the game through experience rather than relying on price representations eliminates potential pyschological complications introduced by such representations in monetized economies (Russo, Krieser, and Miyashita 1975), proving a reference point against which to evaluate the predictions of the theory.

The apparent sequential nature of the pigeons' choices, while standing in contrast to notions of consumers possessing full information and making decisions simultaneously, poses no conflict with respect to observable characteristics of human choice in the marketplace. It takes real time for consumption to occur; hence, real choices of all organisms have a sequential aspect that is commonly ignored in comparative static analysis. Further, differences in "planning horizons" between humans and pigeons must, almost by definition, be one of length rather than existence, unless one is willing to argue that the pigeons' behavior is the result of some elaborate, genetically determined program.

Finally, it must be remembered that whether or not pigeons or humans have consciously thought out their behavior is irrelevant to characterizing that behavior as a solution to a constrained optimization problem. As Samuelson (1947) notes, ". . . it is possible to formulate our conditions of equilibrium as those of an extremum problem, even though it is admittedly not a case of an individual's behaving in a maximizing manner, just as it is often possible in classical dynamics to express the path of a particle as one which maximizes (minimizes) some quantity despite the fact that the particle is obviously not acting consciously or purposively." In a competitive world all organisms must solve the problem that time and energy are limited and that efficient allocation of resources, over which the individual has control, is essential to success. The fact that man is capable of self-awareness and reasoning and is occasionally guided by it in solving mathematical puzzles, for example, does not imply that these factors are either necessary or sufficient to achieving efficient allocation of resources. Nor does this prove that human consumers in the marketplace are guided by these factors, while pigeons in the laboratory are not.

Appendix

Procedural Details

The experiment was conducted in a sound-insulated pigeon test chamber that measured 13 ¼ inches long, 11 ½ inches wide, and 14 inches high. The test

TABLE A1

Subject and Experimental Condition*	No. of Days in Condition	α_F	α_W
Subject 14:			
1	5	60	60
2	12	72	25
3	46	60	60
4	16	34	120
5	14	60	60
Subject 36:			
1	20	60	60
2	19	20	90
3	16	120	13.3
4	56	60	60
Subject 37:			
1	46	60	60
2	19	45	90
3	20	80	20
4	27	60	60
Subject 38:			
1	18	60	60
2	20	48	90
3	15	72	27.7
4	48	60	60
Subject 39:			
1	49	60	60
2	17	48	90
3	15	72	30
4	31	60	60
Subject 46:			
1	14	60	60
2	12	80	28
3	39	60	60
4	25	26	120
5	28	60	60
Subject 50:			
1	8	60	60
2	11	80	20
3	50	60	60
4	19	38	120
5	19	60	60
Subject 85:			
1	16	36	72
2	22	90	20
3	33	60	60

NOTE.—α_F = average time between food payoffs in seconds; α_W = average time between water payoffs in seconds.
* Mean value reported in the text for *S*'s 14 and 50 in condition 1 are over all days in this condition.

chamber had three response keys (2.5 centimeters in diameter, fabricated from clear acrylic plastic) located on the front panel where the food and water reservoirs were located. A single peck on the center response key switched modes of delivery from food to water and vice versa with no changeover delay. This key was transluminated with white light at all times. The other two response keys, located over the food hopper and water reservoirs, were color coded to indicate the delivery mode in operation. Responses on these keys did not affect the contingencies. General chamber illumination was provided by a 7-watt houselight. When commodities were delivered, the response keys were darkened but the houselight remained on. A fan provided ventilation, and constant white masking noise was provided.

Food and water deliveries were programmed by separate variable time tapes. The tapes established a random sequence of interreinforcement times following an exponential distribution under all conditions. The sequence of interreinforcement times repeated itself every 20–30 reinforcements.

Food reinforcements occurred only when in the food mode and consisted of the food hopper coming into reach and the food hopper lights coming on during reinforcement. Water reinforcement occurred only when in the water mode and consisted of pumping water into a reservoir at the rate of 1 milliliter per 15 seconds and operating a response feedback clicker every ½ second. During reinforcement, pecks on the center response key did not result in a mode switch.

Table A1 shows the number of days each condition was in effect and the average absolute time delays (in seconds) between food and water payoffs.

References

Becker, Gary S. "Irrational Behavior and Economic Theory." *J.P.E.* 70, no. 1 (February 1962): 1–13.

———. *Economic Theory.* New York: Knapp, 1971.

Chant, John F. "Irrational Behavior and Economic Theory: A Comment." *J.P.E.* 71, no. 5 (October 1963): 505–10.

Friedman, Milton. *Price Theory.* Chicago: Aldine, 1976.

Hicks, John R. *Value and Capital.* 2d ed. Oxford: Clarendon, 1946.

———. *A Revision of Demand Theory.* London: Oxford Univ. Press, 1956.

Hirshleifer, Jack. "Natural Economy versus Political Economy." *J. Soc. and Biological Structures* 1 (October 1978): 319–37.

Houthakker, Hendrik S., and Taylor, Lester D. *Consumer Demand in the United States: Analyses and Projections.* 2d ed. Cambridge, Mass.: Harvard Univ. Press, 1970.

Kagel, John H., and Battalio, Raymond C. "Token Economy and Animal Models for the Experimental Analysis of Economic Behavior." In *Evaluation of Econometric Models*, edited by Jan Kmenta and James B. Ramsey. New York: Academic Press, 1980.

Kagel, John H.; Battalio, Raymond C.; Green, Leonard; and Rachlin, Howard. "Consumer Demand Theory Applied to Choice Behavior of Rats." In *Limits to Action: The Allocation of Individual Behavior*, edited by John E. R. Staddon. New York: Academic Press, 1980.

Kagel, John H.; Battalio, Raymond C.; Rachlin, Howard; Green, Leonard; Basmann, Robert L.; and Klemm, W. R. "Experimental Studies of Consumer Demand Behavior Using Laboratory Animals." *Econ. Inquiry* 13 (March 1975): 22–38.

Mazur, James E., and Hastie, Reid. "Learning as Accumulation: A Reexamination of the Learning Curve." *Psychological Bull.* 85 (November 1978): 1256–74.

Phlips, Louis. *Applied Consumption Analysis.* New York: American Elsevier, 1974.

Rachlin, Howard. "Economics and Behavioral Psychology." In *Limits to Action: The Allocation of Individual Behavior*, edited by John E. R. Staddon. New York: Academic Press, 1980.

Rapport, David J., and Turner, James E. "Economic Models in Ecology." *Science* 195 (January 28, 1977): 367–73.

Rees, Albert. "Economics." In *International Encyclopedia of the Social Sciences*, edited by David L. Sills. New York: Macmillan and Free Press, 1968.

Russo, J. Edward; Krieser, G.; and Miyashita, S. "An Effective Display of Unit Price Information." *J. Marketing* 39 (April 1975): 11–19.

Samuelson, Paul A. *Foundations of Economic Analysis.* Cambridge, Mass.: Harvard Univ. Press, 1947.

Smith, Vernon L. "Relevance of Laboratory Experiments to Testing Resource Allocation Theory." In *Evaluation of Econometric Models*, edited by Jan Kmenta and James B. Ramsey. New York: Academic Press, 1980.

Stevens, Stanley S. *Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects*, edited by Geraldine Stevens. New York: Wiley, 1975.

Tversky, Amos, and Kahneman, Daniel. "Judgement under Uncertainty: Heuristics and Biases." *Science* 185 (September 27, 1974): 1124–31.

Winer, Ben J. *Statistical Principles in Experimental Design.* 2d ed. New York: McGraw-Hill, 1971.

Taste Change in the United Kingdom, 1900–1955

Steven E. Landsburg

University of Chicago

A search is made for violations of the axiom of revealed preference in aggregate consumption data for the United Kingdom. No such violations are found. This is taken as evidence that tastes remained constant throughout the period under study. The strength of this evidence is estimated within the context of a particular model of taste change.

I. Introduction

This paper investigates the phenomenon of "taste change" in the United Kingdom in the years 1900–1955. In the paragraphs which follow we shall describe a method of searching for evidence of that phenomenon. In Section II we shall carry out several versions of that search and find no such evidence. In Section III we shall develop a general model of taste change in the context of which we shall strengthen the negative result of Section II. Section IV is a further discussion of the model. In Section V we shall discuss some of the implications of our results for economic analysis.

Our basic methodology is similar to that of Houthakker (1963) and Koo (1963). First we choose a set of goods. Then we choose a pair of years—say, for concreteness, 1900 and 1910. We observe what quantity of each good was consumed in 1900 (we shall refer to this set of quantities as the 1900 market basket), and we calculate what the price of this market basket would have been had it been purchased in 1910. If the price is lower than the price of the basket which was actually consumed in 1910, then we conclude that (in 1910) the 1910 basket was considered preferable to the 1900 basket. We repeat this procedure for every possible pair of years in some time period. Then we

[*Journal of Political Economy,* 1981, vol. 89, no. 1]

search for intransitivities—for example, the 1910 basket preferred to the 1900 basket, the 1900 basket preferred to the 1905 basket, and the 1905 basket preferred to the 1910 basket. In other words, we search for violations of the Strong Axiom of Revealed Preference. Such an intransitivity is taken as evidence of a change in tastes.

These methods can be applied either to total national consumption or to per capita consumption. We have done both.

A problem arises when a good is available only in certain years—for example, there appear to have been no raspberries in Britain before 1922. This problem (which was rare) was solved by declaring the good in question to be identical with some other good, the choice of "other good" being dictated partly by aesthetic considerations and partly by the observation that identical goods should have nearly identical price series. (Raspberries were classed with black currants.)

When a good is included in the market basket, it is important to include all close complements and substitutes for that good as well. If, for example, our basket consisted only of automobiles and bicycles, we might discover a preference for more bicycles and fewer automobiles in 1978 than in 1971—reflecting not a taste change but a rise in the price of gasoline. The problem would not arise if gasoline were included in the basket.

II. Evidence

The methods described in Section I were applied to six sets of data, which are described below. No intransitivities were observed in any case.

We first considered total national consumption of nine goods: food, alcoholic beverages, tobacco, clothing, transport and communication, housing, fuel, household durables, and "other." The study covers the years 1900–1955. It is noteworthy that no intransitivities occur throughout this period despite the fact that the definition of the United Kingdom changed with the secession of Ireland in 1920. This may be taken as a small bit of evidence that the tastes of the nation did not change even when the composition of the nation did.

The second study was identical to the first with respect to the time period and the goods considered; it reflected per capita rather than total consumption.

The third and fourth studies were identical to the first and second with the exception that housing was eliminated from consideration. This was because expenditure on housing was defined as the total of rents, rates, and water charges collected, a definition which is clearly objectionable (e.g., it fails to include the opportunity cost of owner-occupied housing).

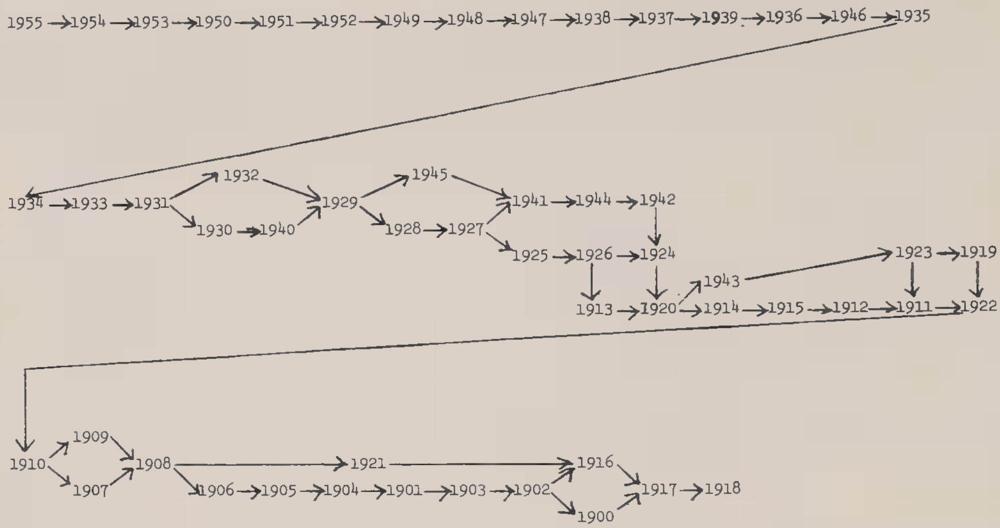


FIG. 1.—Revealed preferences among years for total U.K. consumption of all items, classified into nine categories.

Finally, we broke food, alcohol, and tobacco down into 107 subcategories and studied data on those 107 goods for the years 1920–38. Again this was done both for total and per capita expenditures.

Figure 1 provides an example of the results which were obtained. It depicts the conclusions of the first study described above, concerning total national consumption of nine broadly defined categories of goods. An arrow $A \rightarrow B$ indicates that the consumption of year A was revealed preferred to that of year B. Revealed preferences which can be inferred by transitivity from the ones depicted have been omitted from the figure in the interests of clarity. Analogous charts for the other studies described above are available from the author.

The data for the first four studies mentioned were taken from Mitchell and Deane (1954); for the other two they were taken from Stone and Rowe (1954).

It should be noted that the upward trend in income, particularly toward the end of the period under consideration, would tend to hide changes in taste even if they did occur.[1] However, the tests in the next section will explicitly take this phenomenon into account.

III. Interpretation

Our failure to observe a taste change supports the belief that no taste change occurred; now we ask how strongly that belief can be supported. Upon making certain assumptions and reexamining the data, we shall be able to infer that it is unlikely that the belief is far wrong.

[1] I am indebted to Gary Becker for this observation.

While the model which we shall develop could theoretically be applied to a market with an arbitrary number of goods, practical considerations such as limited computational resources require that we apply it only in the two-good case. For that reason, we shall restrict our detailed description of the model to the two-good case; the results for the n-good case are available on request.

Let X and Y be two goods with prices p_X and p_Y. We shall suppose that the utility function is the "constant elasticity of substitution" function $U(X,Y) = C(AX^u + BY^u)^{1/u}$, where A, B, C, and u are constants and $0 \neq u < 1$. It follows that the elasticity of substitution between X and Y is everywhere equal to $1/(1 - u)$.

We have $dU/dX = AC(AX^u + BY^u)^{(1/u)-1}X^{u-1}$ and $dU/dY = BC(AX^u + BY^u)^{(1/u)-1}Y^{u-1}$. By changing the units in which X and Y are measured we may assume that utility is maximized when one unit of each is consumed, from which we infer that $(dU/dX)/(dU/dY)\big|_{(1,1)} = p_X/p_Y$, that is, $A/B = p_X/p_Y$.

We suppose for the moment that tastes change while relative prices and real income do not. We shall interpret a change in tastes as a shift to a new utility function $V(X,Y)$ which is such that $dV/dX = \alpha(dU/dX)$ and $dV/dY = \beta(dV/dY)$, where α and β are observations of random variables $\hat{\alpha}$ and $\hat{\beta}$. That is, we assume that at the margin, any unit of X or Y will yield α or β times as much utility as it did before. The possibilities $\alpha \leqslant \beta$ and $\beta \leqslant \alpha$ should be equally likely, so we assume that $\hat{\alpha}/\hat{\beta}$ has a median of one. The function satisfying these requirements is $V(X,Y) = C(\alpha AX^u + \beta BY^u)^{1/u}$. This has the same indifference curves as $V(X,Y) = C[(\alpha/\beta)AX^u + BY^u]^{1/u}$, so we may replace α by α/β and β by one. Thus our assumption is that the new utility function has the form $V(X,Y) = C(\alpha AX^u + BY^u)^{1/u}$, where α is an observation of a random variable $\hat{\alpha}$ with a median of one. (The hypothesis that tastes do not change is the hypothesis that $\hat{\alpha}$ has variance zero.)

Now $dV/dX = \alpha AC(A\alpha X^u + BY^u)^{(1/u)-1}X^{u-1}$ and $dV/dY = BC(A\alpha X^u + BY^u)^{(1/u)-1}Y^{u-1}$. In light of the budget constraint, the new levels of consumption must be $1 + t$ units of X and $1 - (A/B)t$ units of Y for some number t. In order to maximize utility, t must satisfy

$$\frac{dV/dX}{dV/dY}\bigg|_{[1+t,1-(A/B)t]} = p_X/p_Y = A/B,$$

that is,

$$\frac{\alpha A(1 + t)^{u-1}}{B[1 - (A/B)t]^{u-1}} = A/B$$

or

$$t = \frac{1 - \alpha^{1/(u-1)}}{(A/B) + \alpha^{1/(u-1)}}.$$

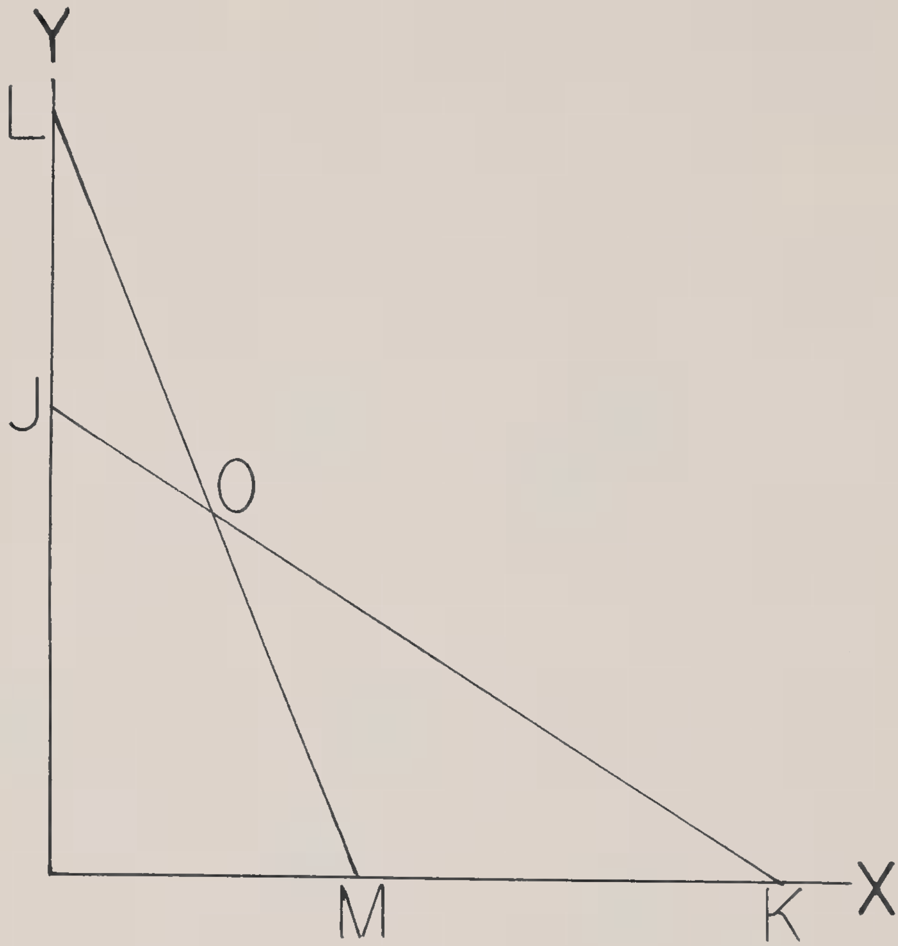


Fig. 2

Now let us suppose that in our second period we have not only a new utility function but a new budget line as well. If the two budget lines fail to cross, then there is no hope of observing a taste change, because no point on the inner budget line can ever be revealed preferred to any point on the outer one. Thus we restrict our attention to the case in which the budget lines *do* cross, as in figure 2. We can rename the goods so that the price of X falls relative to that of Y between the first and second periods; thus we may assume that LM represents the first-period budget line and JK the second-period budget line.

When the basket consumed in the first period is on segment OM and the basket consumed in the second period is on segment OJ, each is revealed preferred to the other and a taste change is observed. An estimate of the probability of this occurrence would yield a measure of the strength of our negative result; however, we shall argue that the hypothesis of taste change can be more strongly rejected through consideration of the "exact opposite" situation—that in which neither

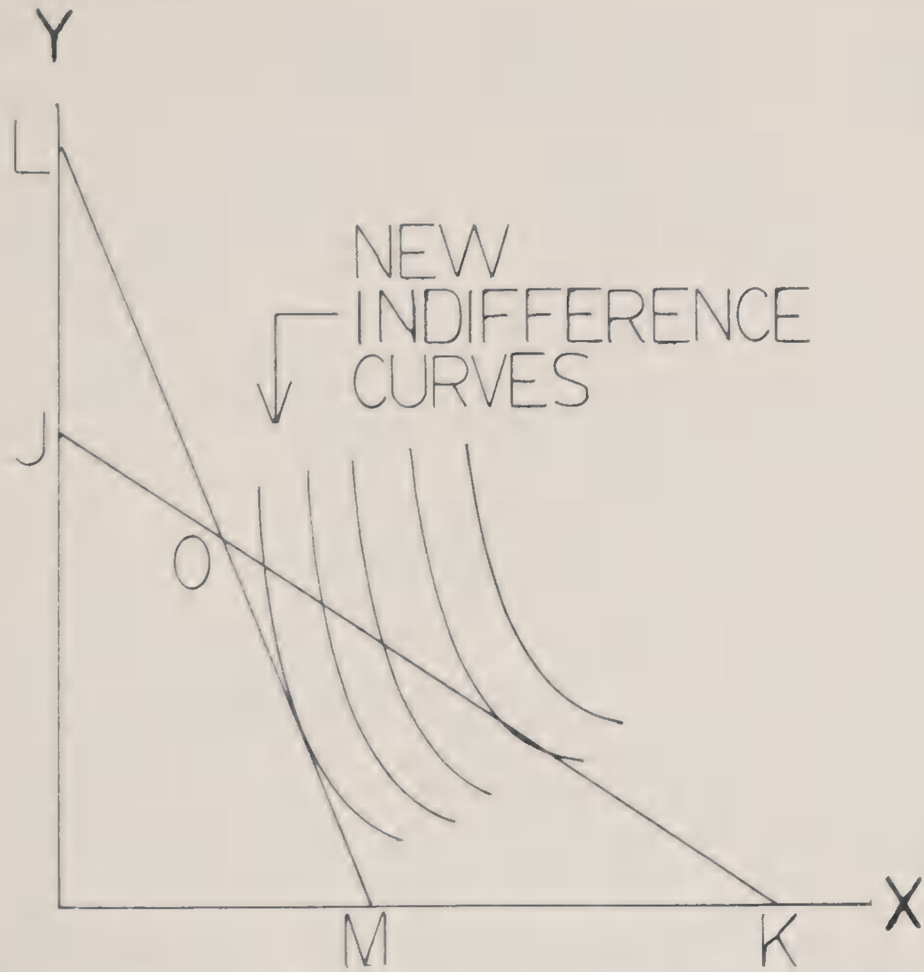


FIG. 3.—If there is an indifference curve tangent to OM, there must be one tangent to OK.

basket is revealed preferred to the other. In terms of figure 2, this situation (which we shall refer to as incomparability of the baskets) arises when the first period's basket is on LO and the second period's basket is on OK.

Now suppose that the first period's basket is observed to be on LO. In keeping with our earlier assumptions, our supposition is that the point (1,1) lies on LO. If the utility function had changed and the budget line had not, the new utility-maximizing market basket would have been at the point $[1 + t, 1 - (A/B)t]$ on line LM. If this point happens to be on OM, then with the new budget line the utility-maximizing market basket will be on OK, as is evident from figure 3.

It follows that if we know that the first period's market basket lies on LO and the point $[1 + t, 1 - (A/B)t]$ lies on OM, we will observe an incomparability.

Denoting the coordinates of 0 by X_0 and Y_0, we see that $[1 + t, 1 - (A/B)t]$ is on OM if and only if $1 + t \geqslant X_0$. Since

$$t = \frac{1 - \alpha^{1/(u-1)}}{(A/B) + \alpha^{1/(u-1)}},$$

this is the case if and only if

$$\frac{1 - \alpha^{1/(u-1)}}{(A/B) + \alpha^{1/(u-1)}} \geqslant X_0 - 1,$$

that is,

$$\frac{1 + (1 - X_0)(A/B)}{X_0} \geqslant \alpha^{1/(u-1)}$$

or—since $A/B = (Y_0 - 1)/(1 - X_0)$—$Y_0/X_0 \geqslant \alpha^{1/(u-1)} = \alpha^{-\eta}$, where η is the elasticity of substitution between X and Y.

Conclusion

If, in figure 3, the first period's market basket is observed to be on LO (i.e., outside the budget constraint of the second period), and if Y_0/X_0 is greater than $\alpha^{-\eta}$, then we will observe an incomparability, provided X and Y are measured in units such that the first period's market basket is at the point (1,1).

Similar considerations show that if the second period's market basket is observed to be on OK and if units are chosen so that its coordinates are (1,1), then an incomparability will be observed provided that $X_0/Y_0 \geqslant \alpha^{-\eta}$.

Remarks

1. Because we shall be applying our model only in the two-good case, we shall have to assume that consumption of all substitutes and complements for the goods in question remains relatively constant (this is so in light of the discussion at the end of Section I). We shall approximate this condition by requiring that the first and second periods be successive years.

2. While the form of the utility function precludes the possibility that $\eta = 1$, we shall assume for ease of exposition that it is very close to one. (In Section IV we will explore the effect of letting η take on other values.) Thus we reinterpret the conclusion as saying that an incomparability will be observed provided Y_0/X_0 or X_0/Y_0, whichever is appropriate, is greater than α^{-1}.

The data which were used to construct figure 1 were reexamined, and values of Y_0/X_0 or X_0/Y_0 were computed for every pair of

successive years and every pair of goods with crossing budget lines. The results, for total and per capita consumption, with and without the questionable housing data, are displayed in table 1 at the end of this section. (In these new studies, as in the earlier ones which involved more than two goods at a time, no intransitivities were observed.)

Under total consumption there are 604 relevant cases (i.e., cases where the budget lines cross). Only nine incomparabilities were observed. More important, there are two cases in which an incomparability would have been observed if α^{-1} had been less than 1.00. If we assume that tastes *do* change (i.e., that $\hat{\alpha}$ is not identically one), then, by our assumption that $\hat{\alpha}$ has median one, α^{-1} must be less than one 50 percent of the time; therefore the probability of our observing an incomparability in one of these two cases is greater than $1 - (1 - 0.5)^2 = 0.75$. Thus our failure to observe an incomparability in either case allows us to reject the hypothesis of taste change with 75 percent confidence. The above remarks apply to total consumption with or without housing data.

Under per capita consumption, there are 673 cases and only three incomparabilities. There are 14 cases in which an incomparability would have been observed if α^{-1} had been less than 0.95. (There are eight such cases if one ignores the housing data.) Among these cases there is one incomparability. If we were to assume that α^{-1} is less than 0.95 even 25 percent of the time, the probability of its being so only once in 14 trials would be only 10.1 percent. Thus we can say with 89.9 percent confidence (63.3 percent confidence if we do not use the housing data) that α^{-1} is not less than 0.95 even as much as 25 percent of the time.

In three cases, none of which involved housing data, X_0/Y_0 or Y_0/X_0 was greater than 0.995. No incomparabilities were observed in any of these cases. Calculating as above, we find that this allows us to reject, with 57.8 percent confidence, the hypothesis that α^{-1} is less than 0.995 more than 25 percent of the time.

Summary of Results

Using total consumption, α^{-1} never differs from one (75 percent confidence). Using per capita consumption, α^{-1} is not less than 0.95 more than 25 percent of the time (89.9 percent confidence, 63.3 percent without housing), and α^{-1} is not less than 0.995 more than 25 percent of the time (57.8 percent confidence). These results say, with some degree of confidence, that $\hat{\alpha}$ seems to have very small variance—and this is just what is meant by the assertion that tastes do not change.

TABLE 1

X_0/Y_0 OR Y_0/X_0	WITH HOUSING DATA		WITHOUT HOUSING DATA	
	No. of Cases	No. of Incomparabilities	No. of Cases	No. of Incomparabilities
Total consumption:				
.00–.049	54	0	46	0
.05–.099	40	0	31	0
.10–.149	43	0	32	0
.15–.199	39	0	27	0
.20–.249	54	0	43	0
.25–.299	45	0	33	0
.30–.349	34	0	26	0
.35–.399	30	0	21	0
.40–.449	22	0	16	0
.45–.499	28	0	22	0
.50–.549	33	0	27	0
.55–.599	20	0	15	0
.60–.649	28	0	21	0
.65–.699	23	0	14	0
.70–.749	25	0	20	0
.75–.799	21	0	17	0
.80–.849	20	1	13	1
.85–.899	13	0	11	0
.90–.949	15	0	8	0
.95–.999	15	8	10	7
1.00	2	0	2	0
Total	604	9	455	8
Per capita consumption:				
.00–.049	62	0	47	0
.05–.099	44	0	38	0
.10–.149	52	0	40	0
.15–.199	56	0	37	0
.20–.249	51	0	39	0
.25–.299	42	0	30	0
.30–.349	39	0	22	0
.35–.399	32	0	25	0
.40–.449	34	0	25	0
.45–.499	32	0	23	0
.50–.549	30	0	25	0
.55–.599	26	0	21	0
.60–.649	29	0	19	0
.65–.699	29	0	21	0
.70–.749	22	0	17	0
.75–.799	21	0	11	0
.80–.849	24	0	20	0
.85–.899	15	0	9	0
.90–.949	19	2	13	1
.95–.999	13	1	7	1
1.00	1	0	1	0
Total	673	3	500	2

It is important to note that the results listed above were derived without any reference to the fact that in all of our studies we never observed a taste change (the phrase "all of our studies" refers both to the set of studies described in Section II and to the 1,277 separate studies described in this section). This striking fact, which is really the main result of this paper, surely mitigates our disappointingly low levels of confidence, but I know of no way in which to quantify the extent to which it does so.[2]

IV. Discussion

Like most models, ours is laden with assumptions with which a reasonable man might take issue. Here we reconsider some of the points which are most open to question.

1. *The Form of the Utility Function*

It is certainly the case that our choice is arbitrary. However, utility functions of this form are often assumed in economics, and our results do indicate that we should be wary of models which assume such utility functions and also allow for changes in taste. Thus, even if we were to grant that our utility function is not an accurate reflection of economic reality, we would still be able to draw conclusions about the consistency of certain economic models, and this is a topic which should be of concern to economists.

It should also be noted that, after some effort, I have not been able to find a "reasonable" utility function which yields results substantially different than does the one adopted here.

[2] One way to begin to quantify it is the following: *Let* p_X^i and p_Y^i be the prices of X and Y in the ith year and let

$$D = \frac{p_X^2 p_Y^1}{p_X^1 p_Y^2}$$

(so that D is always less than one, because the goods were renamed so that the relative price of X fell). Then it can be shown that a taste change will be observed if and only if $\alpha < (X_0/Y_0)D$ or $\alpha < (Y_0/X_0)D$ (depending on whether the first year's basket is inside the second year's budget line or vice versa). Now we could proceed as in the body of the paper, measuring values for $(X_0/Y_0)D$ and $(Y_0/X_0)D$ instead of X_0/Y_0 and Y_0/X_0. This procedure—which amounts to measuring the strength of our failure to observe a taste change by measuring the probability that a taste change could occur and go unobserved—seems more straightforward than the one adopted in the text, but when applied to the same data it yields numerical results which are substantially weaker. Given the assumptions of our model, this means that even if tastes changed substantially, our failure to observe a taste change in any of the 1,277 cases of Section III would not be too surprising. However, this calculation ignores all of the findings of Section II, which appear to be quite strong. I do not know the best way to quantify their strength—a naive generalization of the current model would require additional assumptions and very extensive computational resources.

2. *The Exogeneity of Price Changes*

On the surface this is a very damaging objection. If prices and tastes do not change independently, as we have assumed, than a considerably subtler analysis of the problem is required. However, reasoning as in (1) above, we argue that the fact that many economic models do assume exogenous price changes implies that we can at least draw a conclusion about those models. The relevant question for economists is not "Do tastes change?" but "Are models which assume taste changes consistent with reality?" The latter is the question which we have really been addressing; the answer, for a large class of models, appears to be "no."

3. *The Form of the Taste Change*

All that can really be said here is that the form which we have assumed seems reasonable and that other reasonable forms yield similar results. If, for example, we had assumed that the new utility function was given by $V(X,Y) = U(\alpha X,Y)$, rather than being determined by the requirement that $dV/dX = \alpha(dU/dX)$, then our conclusion would have been exactly the same except that $\alpha^{1-\eta}$ would have appeared instead of $\alpha^{-\eta}$.

4. *The Assumption That η Is Close to One*

Our calculation of the confidence with which we can say that α never differs from one (i.e., that tastes do not change at all) is unaffected by this assumption: It would remain the same for any value of η. With regard to our calculation of the confidence with which we can say that α is not less than some given value more than a given percentage of the time, our result would be stronger for larger η and weaker for smaller η.

5. *The Assumption That α Is Drawn Independently Each Time*

This assumption was essential when we calculated confidence levels in Section III. Our first calculation was that we could say with 75 percent confidence that α never differs from one. The two cases involved in this calculation were transport and communication versus tobacco for 1907–8 and fuel versus alcoholic beverages for 1937–38. The 30-year separation, together with the fact that entirely different goods are involved, seems to justify the assumption that α was drawn independently in these two cases.

Our strongest numerical result involved the confidence with which we could say that α was not less than 0.95 more than 25 percent of the

time; it was calculated to be 89.9 percent on the basis of 14 cases. The 14 cases involved are listed in the unnumbered table below.

1. Housing vs. food	1913–14
2. Housing vs. alcoholic beverages	1942–43
3. Housing vs. tobacco	1916–17
4. Housing vs. tobacco	1943–44
5. Housing vs. fuel	1951–52
6. Household durables vs. tobacco	1922–23
7. Household durables vs. fuel	1927–28
8. Household durables vs. fuel	1929–30
9. Household durables vs. transport and communication	1905–6
10. Household durables vs. transport and communication	1909–10
11. Alcoholic beverages vs. other	1909–10
12. Household durables vs. other	1904–5
13. Housing vs. other	1902–3
14. Fuel vs. other	1901–2

(Case 14 was the one in which an incomparability was observed.) The reader is invited to form his own opinion as to how many of these cases are really independent. If, for example, he discards cases 8 and 10 on the grounds that they are too closely tied to cases 7 and 9, then we have 12 independent cases instead of 14 and our confidence drops from 89.9 percent to 84.2 percent.

V. Conclusions

The main result of this paper is that no taste change was observed in any of our studies. These included studies of total consumption and of per capita consumption, studies of broadly defined categories of goods and of very narrowly defined goods, studies of exhaustive and nonexhaustive lists of goods, and studies of as many as 107 and as few as two different goods. The two-good studies alone presented 1,980 separate cases; 1,277 of these had crossing budget lines and hence the potential for an observed taste change. None of this potential was realized. In Section III we calculated levels of confidence with which we could say that tastes did not change by more than certain specified amounts; these calculations were independent of the fact that no taste change was observed and therefore greatly understate the strength of our results.

The morals are the obvious ones: In economics, we should be wary of models which include variables for taste change, and outside economics we should be skeptical of warnings about the "wild vicissitudes of taste" and the constant danger of having our vicissitudes manipulated. With regard to this last point, it is particularly interesting to note that our study covers the period in which the electronic media came into being and into flower.

It is easy to postulate specific forms of taste change among the goods which we have studied, and it is easy to postulate forms which our methods could not have observed. It is considerably harder to postulate such forms which satisfy the additional restriction that they be reasonable. I do not believe that is possible, at least not without violating exogeneity of prices. If this is correct, then our results have strong implications for all price-theoretic models in which price changes are exogenous.

References

Houthakker, Hendrik S. "Some Problems in the International Comparison of Consumption Patterns." In *Les Besoins de biens de consommation*, an international colloquium of the Centre National de la Recherche Scientifique. Grenoble: Centre National de la Recherche Scientifique, 1963.

Koo, Anthony Y. C. "An Empirical Test of Revealed Preference Theory." *Econometrica* 31 (October 1963): 646–64.

Mitchell, Brian R., and Deane, Phyllis. *Abstract of British Historical Statistics.* Cambridge: Cambridge Univ. Press, 1954.

Stone, J. Richard, and Rowe, D. A. *The Measurement of Consumers' Expenditure and Behaviour in the United Kingdom, 1920–38.* Cambridge: Cambridge Univ. Press, 1954.

The Determinants of Tariff and Nontariff Trade Restrictions in the United States

Edward John Ray

Ohio State University

This paper develops and tests a simple model for the determination of tariff and nontariff barriers to trade across industries within the United States, using 1970 trade data. We find that nontariff trade restrictions have supplemented tariff protection in the United States. Both tariff and nontariff trade restrictions are biased toward industries in which the United States has an apparent comparative disadvantage in world trade and away from industries in which consumer welfare losses from protection would be great. We also find substantial evidence that tariff and nontariff trade restrictions predominate in industries with very different market characteristics.

The purpose of this study is to synthesize a number of ideas expressed in earlier work into a simple analytical framework to explain the structure of tariff and nontariff barriers to trade across industries in the United States. Our central premise is that, subject to political constraints, trade restrictions are consistent with profit maximization across industries. Using International Trade Commission data for tariff and nontariff trade restrictions and industry characteristics for 225 U.S. manufacturing industries in 1970, we find that such an approach has significant explanatory power.

The resulting empirical analysis is unique in a number of important respects. First, we provide separate reduced form estimates of tariff

The author is grateful to Betsy Radigan, Steve Cook, Ed Honton, and Jim Brown for aid in generating the Tobit and Probit results. Professor Tetsunori Koizumi, Jacob Frenkel, Howard Marvel, Donald Parsons, Jerry Thursby, Marie Thursby, Thomas Wolf, and, most particularly, J. David Richardson provided numerous helpful comments on earlier drafts of this paper.

[*Journal of Political Economy*, 1981, vol. 89, no. 1]

and nontariff trade restrictions in the United States. Second, by extending our analysis to include the estimation of a simultaneous model for the determination of tariff and nontariff trade restrictions, we provide empirical support for our hypothesis that nontariff trade restrictions have been utilized in part to compensate for internationally agreed-upon tariff reductions in the post–World War II period.

The third important finding is that, while tariff and nontariff trade restrictions co-exist in industries with some common characteristics, there are some distinct differences in industry characteristics that are more strongly associated with one rather than the other alternative form of protection. Specifically, we find that both tariff and nontariff trade restrictions are biased toward industries in which the United States is at a comparative disadvantage in world trade, and away from industries in which losses in consumer welfare from protection would be particularly large. But there are some real differences too. Tariffs are biased toward industries which are low-skill intensive, away from industries that are capital intensive, and unrelated to product heterogeneity, the concentration of production, and the geographical dispersion of domestic production facilities. In contrast, nontariff trade restrictions are concentrated in industries producing fairly homogeneous products using relatively capital-intensive techniques of production that, at the same time, are not intensive users of low-skilled labor. In addition, nontariff trade restrictions are found predominantly in industries in which production is less concentrated and in which production facilities are distributed across regions of the United States in a fashion consistent with the distribution of population and, therefore, voting power in Congress.

I. The Political Economy of Trade Restrictions

The general approach of this paper is that the structure of tariff and/or nontariff trade restrictions across industries is consistent with the simple joint maximization of industry profits subject to political constraints. In effect, we assume that individual industry characteristics influence the profitability of trade restrictions and that political factors similar to those discussed by Cheh (1974, 1976), Pincus (1975), Baldwin (1976*a*), Caves (1976), Caves and Jones (1977), Helleiner (1977), Stone (1978), and others are critical in determining which industries are most successful in obtaining restrictions on imports of competitive products.

The primary information to emerge from a simple analysis of the relationship between industry rents and protection is that the profitability of trade restrictions should normally be negatively related to the absolute value of the own price elasticity of demand for the product

and positively related to the foreign elasticity of supply of the product.[1] In effect, the more responsive consumers are to price increases induced by trade restrictions and the less responsive foreign suppliers are to changes in the international price of the good, the less profitable any given level of protection will be.

Furthermore, a discussion of the profitability of protection assumes that imports represent a reasonable threat to sales in the United States by domestic producers. If concern for industry profits is important in explaining the structure of trade restrictions, we would expect trade barriers to be biased away from industries in which the United States has a comparative advantage in international trade because imports in such industries would be expected to be inconsequential and restricting them would provide little benefit.[2] We included a number of variables that previous studies have indicated are associated with U.S. comparative advantage or disadvantage in trade with the simple expectation that trade restrictions should be found predominantly in product lines in which the United States is at a comparative disadvantage vis-à-vis the rest of the world.

Obviously, if there are positive terms of trade effects associated with protection sufficient to offset the deadweight consumption and production costs created by trade barriers, the government can generally increase aggregate social welfare and support for itself by restricting trade. But substantial terms of trade effects are not likely to exist for more than a handful of industries at present levels of protection. In general, increased protection would be expected to promote industry profits at the expense of the general public. In a broad sense, profits should be thought of as the sum of rents to entrepreneurs and to workers with firm- or industry-specific human capital and/or a strong union. The costs to the general public include artificially high product

[1] Assuming that protection is less than prohibitive, that foreign supply depends only on the foreign price, and that the domestic market clearing price, P_D, is positively related to the degree of trade protection given, we can generate the following relationship between industry profits, Π, and trade restrictions:

$$\frac{\partial \Pi}{\partial \tau} = \left\{ S_D + \left(\frac{MC_D}{P_D} - 1 \right) \left[D\eta + S_F \epsilon_F \left(1 - P_F \frac{\partial \tau}{\partial P_D} \right) \right] \right\} \frac{\partial P_D}{\partial \tau}$$

where S_D represents domestic supply, MC_D is the domestic marginal cost of production, D is aggregate domestic demand, η = the absolute value of the own price elasticity of demand for the product, S_F, ϵ_F, and P_F represent the foreign supply, foreign supply elasticity, and foreign price of the product, respectively, and τ represents the tariff or tariff equivalent rate. Details for this derivation are available upon request from the author.

[2] There is, of course, the interesting possibility that the United States has a comparative advantage in some product and that trade restrictions are used to price discriminate between the domestic and foreign market to maximize monopoly rents. That such a phenomenon could be systematic and significant across the 225 manufacturing industries we have data for is doubtful.

prices, misallocation of productive resources, and waste in terms of administrative costs of implementing and maintaining protectionist programs.

In order to rationalize the existence of trade restrictions for numerous commodities in the United States, we assume that, up to some limiting degree of trade restraint, industry interest groups are more responsive to trade policy than consumers are.[3] One simple explanation of such an asymmetry in response to trade policy by industry interests and by consumers stems from the fact that information about market conditions is costly and imperfect and must be updated as market conditions change. To the extent that producers are less numerous than consumers and their well being (wealth) is more directly affected than the wealth of individual consumers by trade policy, they will invest more resources in keeping informed about and be more sensitive to trade policy changes affecting their industry. Clearly, the reverse could be true for industries in which buyers are more concentrated than sellers and the wealth effects are greater per consumer than they are per producer. But such cases are exceptional. This asymmetry in responsiveness does not imply that seller concentration need be positively related to protection. In fact, in his study of the structure of tariffs in Canada, Caves (1976) suggested that concentration could be negatively related to protection. Presumably, the argument would be that the spoils of protection will buy more votes the more widely they are dispersed. In each industry there may exist some critical value of seller concentration relative to buyer concentration, determined by the cost of market information which is required for the industry to obtain protection. And among protected industries there may be a negative association between protection and concentration, CONR. In short, the impact of seller concentration on the existence and extent of trade restrictions is ambiguous.

So far we have argued that producers and workers in an industry have a greater economic incentive on average than consumers in general do to acquire costly information about the impact of trade restrictions on their prices, profits, and wages and in continuing to invest in such information gathering as market conditions change and old information decays or becomes less relevant. However, constant changes in trade restrictions could easily attract media attention and thereby provide consumers with relatively cheap and current infor-

[3] Apart from the observation that nonzero trade restrictions exist for industries that enjoy no obvious terms of trade effects, we provide no direct test of our assumption that industry interests are more responsive than consumer interests to changes in trade policy. However, the notion that industry interests are weighted more heavily than consumer interests in determining the structure of protection is consistent with earlier empirical work by Ray (1974), Baldwin (1976*b*), and Caves (1976).

mation regarding the impact that proposed changes in trade limitations would have on their welfare. Therefore, pressure to alter trade restrictions may follow a fairly discrete time path in order to permit industry interest groups to capitalize on the high cost and low quality of information available to consumers.[4] Maintaining a given quota in the presence of expanding demand or foreign supply would be comparable to imposing higher and higher explicit tariffs over time. Conversely, if demand for a product is declining and/or foreign supply is shrinking, a quota is comparable to a declining explicit tariff. Therefore, assuming some stickiness in policy changes, quotas or other nontariff trade restrictions would have some relative advantage in expanding markets while tariffs would be somewhat advantageous for protecting declining product markets.[5] In our empirical analysis we will attempt to test the relative explanatory power of dynamic demand, ΔD, and supply, ΔS, conditions in tariff and nontariff trade restriction regressions.

Finally, simple demand and supply analysis can be used to demonstrate that the deadweight loss associated with a given trade restriction will be positively related to the absolute value of the own price elasticity of demand for the product, η, and positively related to the domestic elasticity of supply, ϵ_D. To the extent that the government is strongly concerned with the overall welfare effects of trade restrictions, tariff and nontariff trade barriers would be negatively related to the absolute value of the own price elasticity of demand for the product and negatively related to the domestic elasticity of supply for the product.

We can summarize the discussion to this point by indicating the tariff and nontariff trade barrier regressions to be estimated, with the expected sign of the coefficient specified below each variable:

$$\begin{array}{l} \tau = F(\eta, \epsilon_F, \epsilon_D, X_1, \ldots, X_n, \text{CONR}, \Delta\text{D}, \Delta\text{S}) \\ \qquad\quad -, \ ?, \ -, \ -, \ldots, \ -, \quad ?, \quad +, \ + \end{array} \tag{1}$$

and

$$\begin{array}{l} N = G(\eta, \epsilon_F, \epsilon_D, X_1, \ldots, X_n, \text{CONR}, \Delta\text{D}, \Delta\text{S}, \tau) \\ \qquad\quad -, \ ?, \ -, \ -, \ldots, \ -, \quad ?, \quad +, \ +, \ + \end{array} \tag{2}$$

where N represents a measure of nontariff trade restrictions, the ambiguous sign on the elasticity of foreign supply, ϵ_F, reflects the fact that the sign on the foreign supply elasticity would only be positive in

[4] Similar views regarding the relative merits of tariffs and quotas in the context of dynamic market changes have been expressed by Kreinin (1970), Fishelson and Flatters (1975), Caves and Jones (1977), and others.

[5] Cheh (1974) found some evidence that tariff cuts during the Kennedy Round were less vigorously applied to declining industries in the United States.

the noncompetitive case, ϵ_D represents the domestic supply elasticity, $X_1, \ldots, X_n$ are alternative measures of U.S. comparative advantage, CONR is a measure of seller concentration, and ΔD and ΔS represent the expected percentage growth in demand and supply at the time that trade restrictions are set (both are expected to be more positively significant in explaining nontariff than tariff protection).

In the empirical section we treat the implementation of tariff and nontariff trade barriers as a sequential process. Tariffs predate nontariff barriers in virtually every case, and nontariff trade restrictions are estimated as functions of both tariffs and of the political and economic factors we identify as important explanatory variables. Later we test for this sequential effect and find evidence that it does exist.

II. Empirical Results

The empirical results discussed in this section were generated from data made available by the U.S. International Trade Commission in 1975. More precise definitions of the variables used in the regressions are given in the Appendix. The observations consisted of 225 four-digit manufacturing industries in the United States in 1970 and are specified in a separate appendix available upon request.

Alternative forms for tariff regressions were estimated and are presented in table 1, while estimates of alternative specifications for nontariff trade barrier regressions are presented in tables 2 and 3. For regressions (1.1)–(1.3) in table 1, the dependent variable used was a weighted-average nominal tariff measure for each industry in 1970. In effect, the four-digit tariff rates were calculated by weighting the tariffs of less aggregated components by their import shares within the classification for the given four-digit industry. Regressions (1.4)–(1.6) differ from the earlier ones only in terms of the specification of the dependent variable. In regressions (1.4)–(1.6) the dependent variable is the four-digit industry nominal tariff rate calculated as the simple average of component nominal tariff rates.

The dependent variable in table 2 is an Index of the Incidence of Nontariff Barriers in the U.S., 1970, constructed by the U.S. Tariff Commission.[6] In effect, the index measures the comprehensiveness of

[6] The index was constructed by the U.S. Tariff Commission and is explained in *Trade Barriers*, Report to the Committee on Finance of the U.S. Senate, Part 2, pp. 160–72, Washington, D.C., April 1974. The quantitative restrictions included and the weighting scheme used were as follows: Bilateral Quota (0.91); Global Quota (1.96); Quota (unspecified) (1.36); Prohibited Imports (embargoes) (1.36); State Trading (0.91); Automatic Licensing (0.45); Liberal Licensing (0.45); Discretionary Licensing (0.91); Licensing (unspecified) (0.91); Minimum Price System (1.36); Seasonal Restriction (0.91); Restriction (unspecified) (0.91); Export Restraint (VERs) (1.36); Suspended

TABLE 1*

U.S. Tariff Regressions, 1970 (OLSQ)†

Independent Variables	Dependent Variables					
	U.S. Weighted-Average Tariffs: 1970			U.S. Simple Average Tariffs: 1970		
	(1.1)	(1.2)	(1.3)	(1.4)	(1.5)	(1.6)
Constant	17.96	16.22	9.94	18.62	17.09	14.95
	(14.15)	(9.74)	(4.32)	(14.14)	(9.76)	(6.05)
FTDNMOS	...	$-.6 \times 10^{-3}$	$.4 \times 10^{-3}$	...	$.3 \times 10^{-3}$	$.9 \times 10^{-3}$
		(.27)	(.21)		(.15)	(.39)
CONR4L	.01	.01	.02	.02	.02	.02
	(.74)	(.59)	(1.06)	(1.46)	(1.31)	(1.23)
SKILLD	−31.85	−30.96	−33.35	−31.46	−30.78	−32.37
	(9.26)	(8.84)	(10.00)	(8.82)	(8.36)	(9.04)
SEINRD	...	7.95	10.29	...	5.90	7.32
		(2.13)	(2.86)		(1.51)	(1.90)
ESCAL(67)	−6.69	−6.57	−4.57	−4.11	−4.06	−3.30
	(2.80)	(2.79)	(1.94)	(1.66)	(1.64)	(1.31)
LABINT	...	...	10.73	...	...	3.36
			(3.63)			(1.06)
KLRA	...	$-.3 \times 10^{-4}$	...	...	$-.2 \times 10^{-4}$	...
		(2.23)			(1.59)	
WDAVPD	.25	−.01	−.18	.14	−.05	−.07
	(.64)	(.04)	(.46)	(.34)	(.11)	(.17)
ΔIMP	.04	.02	$-.45 \times 10^{-2}$	.05	.03	.03
	(.47)	(.25)	(.06)	(.53)	(.39)	(.37)
ΔCON	−.70	−1.11	−.08	−.83	−1.10	−.74
	(.60)	(.96)	(.07)	(.69)	(.90)	(.59)
R^2	.33	.37	.39	.30	.32	.31
F-statistic‡	18.11	13.88	15.29	15.46	11.11	10.88

* Absolute values of t-statistics appear in parentheses. For samples of this size t-statistics greater than 1.65 are significant at the 5% level and t-statistics greater than 2.33 are significant at the 1% level. (For a two-tailed test the values of the t-statistic for significance at the 5% and 1% levels, respectively, are 1.96 and 2.576.)

† OLSQ = Ordinary Least Squares. The weighted-average tariff and simple average tariff regression results were not estimated using Tobit since only 5 of the 225 observations were at the limit value of zero for the weighted tariffs and only 1 observation was at the limit value for the simple tariffs.

‡ For $F(6,218)$ an F-statistic greater than 2.80 is in the upper 1% of the distribution. For $F(9,215)$ an F-statistic greater than 2.41 is in the upper 1% of the distribution.

TABLE 2

U.S. Nontariff Trade Restrictions (Tobit)

Independent Variables	Dependent Variable: Index of the Incidence of Nontariff Barriers in the U.S.: 1970 (2.1)	(2.2)	(2.3)	(2.4)	(2.5)	(2.6)
Constant	1.91	.26	6.68	2.12	.38	6.16
	(1.02)	(.11)	(2.43)	(1.11)	(.16)	(2.13)
FTDNMOS	...	$.14 \times 10^{-2}$	$.1 \times 10^{-4}$	...	$.12 \times 10^{-2}$	$-.1 \times 10^{-4}$
		(.54)	(.38)		(.46)	(.03)
CONR4L	−.04	−.05	−.06	−.04	−.05	−.06
	(1.92)	(2.36)	(2.86)	(1.98)	(2.44)	(2.74)
SKILLD	−5.05	−6.78	−2.42	−6.26	−8.16	−6.04
	(1.13)	(1.49)	(.53)	(1.42)	(1.82)	(1.37)
SEINRD	...	4.99	2.14	...	5.95	4.06
		(1.08)	(.47)		(1.29)	(.89)
ESCAL(67)	2.44	2.78	.34	1.61	2.01	−.31
	(.88)	(1.00)	(.12)	(.59)	(.73)	(.11)
LABINT	...	...	−12.69	...	...	−10.05
			(3.76)			(3.07)
KLRA	...	$.1 \times 10^{-5}$	...	...	$.1 \times 10^{-5}$	...
		(2.14)			(2.02)	
WDAVPD	−3.93	−4.004	−3.75	−3.76	−3.87	−3.59
	(3.62)	(3.63)	(3.36)	(3.49)	(3.53)	(3.26)
ΔIMP	−.07	−.05	−.03	−.08	−.06	−.05
	(.75)	(.56)	(.34)	(.83)	(.64)	(.50)
ΔCON	1.40	1.34	.09	1.45	1.33	.24
	(1.01)	(.95)	(.07)	(1.04)	(.93)	(.17)
USWTTF	.20	.20	.26	...	...	...
	(2.97)	(2.99)	(3.72)			
USSMPTAR	...	...	...	.18	.18	.19
				(2.70)	(2.70)	(2.84)

Note.—158 of the 225 observations had the limit value of zero.

nontariff protection in an industry. Fifteen types of nontariff trade restrictions were considered and assigned weights reflecting their relative effectiveness in limiting imports. Then the extent of nontariff protection given to each industry was calculated as the percentage of actual to potential protection to be derived from all 15 nontariff

---

Import Restriction (0.91); Mixing Regulations (0.91). The weights on the right were determined by assigning each restriction a number from 1 to 3, high or low, depending on its degree of restrictiveness. The average of those numbers was divided into the assigned number for each restriction, resulting in the weights shown. Index values for industry categories were then obtained as follows: For any given commodity, the sum of its specific weights divided by the sum of weights for the 15 categories (11.35) expressed in percentage terms yielded the percentage of maximum nontariff restraint of trade given to that industry. Total imports of major trading countries were used as weights to aggregate the basic product category data. Such a weighting means that restrictions in a heavily traded category were counted more than restrictions in lightly (world) traded categories. In cases in which the listed restrictions were known to apply to only part of a given product category they were arbitrarily counted only half as much as restrictions known to be applicable to an entire category.

TABLE 3

U.S. NONTARIFF TRADE RESTRICTIONS (Probit)

INDEPENDENT VARIABLES	DEPENDENT VARIABLE: Dummy Variable for Nontariff Barriers in the U.S.: 1970 (3.1)	(3.2)	(3.3)	(3.4)	(3.5)	(3.6)
Constant	.20	−.26	2.35	.36	−.09	2.36
	(.41)	(.41)	(2.75)	(.72)	(.15)	(2.77)
FTDNMOS	. . .	$.3 \times 10^{-3}$	$.1 \times 10^{-2}$	. . .	$-.3 \times 10^{-3}$	$.1 \times 10^{-2}$
		(.34)	(1.20)		(.45)	(1.28)
CONR4L	−.01	−.01	−.02	$-.7 \times 10^{-2}$	−.01	−.02
	(1.38)	(2.36)	(3.10)	(1.46)	(2.41)	(2.96)
SKILLD	−1.79	−2.77	−1.47	−2.13	−3.14	−2.51
	(1.52)	(2.21)	(1.13)	(1.84)	(2.55)	(2.01)
SEINRD	. . .	2.68	1.86	. . .	2.84	2.22
		(2.21)	(1.46)		(2.37)	(1.80)
ESCAL(67)	−.82	−.59	−1.73	−.99	−.79	−1.82
	(1.09)	(.74)	(1.96)	(1.34)	(1.00)	(2.14)
LABINT	. . .	. . .	−4.79	. . .	. . .	−4.01
			(4.75)			(4.22)
KLRA	. . .	$.1 \times 10^{-4}$	. . .	. . .	$.1 \times 10^{-4}$	. . .
		(2.95)			(2.78)	
WDAVPD	−.41	−.47	−.40	−.39	−.46	−.38
	(2.53)	(2.78)	(2.22)	(2.39)	(2.68)	(2.16)
ΔIMP	−.27	−2.29	−.02	−.03	−.02	−.02
	(1.05)	(0.87)	(.60)	(1.07)	(.92)	(.69)
ΔCON	.52	.48	.04	.50	.44	.02
	(1.46)	(1.29)	(.11)	(1.41)	(1.19)	(.05)
USWTTF	.06	.06	.08	. . .	. . .	. . .
	(2.77)	(2.80)	(3.53)			
USSMPTAR	. . .	. . .	. . .	.05	.05	.05
				(2.49)	(2.51)	(2.77)
Pseudo R^2	.25	.31	.39	.24	.30	.35
Likelihood Ratio Test*	44.29	56.42	73.03	42.29	54.25	65.85

* For eqq. (3.1) and (3.4), a value above 18.50 for the likelihood ratio test is in the upper 1% of the distribution. For the remaining equations, a value above 23.2 for the likelihood ratio test is in the upper 1% of the distribution.

restrictions. Industries with high percentage index values are then viewed as having received more comprehensive nontariff protection than industries with low percentage index values. The dependent variable in table 3 is simply a dummy variable which equals 1 if the index for nontariff barriers is positive and zero otherwise. In every other respect the equations estimated in tables 2 and 3 are the same.

There are three different specifications for each of the tariff relationships and six different specifications for each of the nontariff relationships because we have taken an eclectic approach to the empirical explanation of U.S. comparative advantage. Within each subset of equations the alternative regressions differ only in their specifica-

tion of the $X_1, \ldots, X_n$ variables alluded to in Section I and in the specification of the tariff variable in the nontariff trade restriction regressions. The various specifications of variables explaining U.S. comparative advantage are derivative from previous empirical works, including Stern (1964), Hufbauer (1970), Baldwin (1971), Caves (1976), Cheh (1976), and Stone (1978). The tariff regressions were estimated using ordinary least squares. Since only 67 of the 225 industries had nonzero, nontariff trade restrictions, the nontariff trade barrier index equations and the nontariff trade barrier dummy equations were estimated using Tobit and Probit estimating techniques, respectively.

International trade agreements and the decline in the importance of tariffs as a source of federal revenues have tended to induce the substitution of nontariff trade restrictions for tariffs in protected industries. The positive significant coefficient on the tariff variable in each of the nontariff trade restriction dummy equations indicates that nontariff trade restrictions are concentrated in industries with high nominal tariffs. Whether those nontariff trade restrictions were induced by tariff ceilings in highly protected industries, or not, is not clear. What is clear is that industries with the economic need and political clout to obtain high rates of tariff protection are also successful in obtaining nontariff trade restrictions. The use of nontariff barriers to trade in the United States apparently results in a greater variation in trade restraint across industries than one would suspect from looking at nominal tariff rates alone.

The positive and significant coefficient on the tariff variable in each of the nontariff trade restriction equations in table 2 indicates that high tariffs and high rates of nontariff trade restrictions go hand in hand. In effect, nominal tariffs alone both understate the extent of variation in protection across industries and systematically understate the extent to which high tariff industries are being protected.

To the extent that tariffs have been limited by domestic and international political considerations as well as industry preferences for nontariff trade restrictions, we would expect the economic and political factors described in Section I to maintain significant explanatory power in the nontariff trade barrier regressions. The empirical results discussed below are consistent with that expectation.

Hufbauer (1970), Baldwin (1971), and others have found that significant correlations exist between the ability of the United States to export products and the skill intensity, research and development intensity, and labor intensity of production. Our objective is not to replicate previous work but, rather, to introduce variables that have been employed as proxies for U.S. comparative advantage in previous work as control variables that should enter the tariff and nontariff

trade restriction equations with negative and significant coefficients. Confirmation of that expectation would be consistent with our hypothesis that protection is provided partly in response to industry pressure to reduce competition from foreign suppliers in the domestic market.

The measures used as controls for comparative advantage include the first trading date, FTDNMOS; the percentage of scientists and engineers in research and development, SEINRD; and the percentage of skilled workers in the workforce, SKILLD.[7]

Focusing on the results in tables 1 and 3,[8] only the skill intensity measure has the expected negative and significant coefficient in the tariff and nontariff regressions. Numerous studies of trade flows have found that the United States imports relatively capital-intensive goods and goods which use low-skilled labor suggesting that protection would be biased toward industries with both characteristics. However, we find that tariffs are negatively related to the capital/labor ratio, KLRA, and positively related to labor intensity, LABINT, holding the skill coefficient constant, while nontariff trade restrictions are positively related to the capital/labor ratio and negatively related to the labor intensity of production.

One could speculate that, since recent additions to trade restrictions have been primarily in the form of nontariff barriers to trade, the negative coefficient on LABINT reflects an inability of nonunionized, low-skilled labor to obtain protection from imports while middle-skilled, unionized workers and producers in industries with heavy capital components have succeeded in obtaining nontariff trade restrictions. Then the reverse signs on those variables in the tariff regressions would have to be attributed to the upper bounds on tariff changes that have been imposed in the Post–World War II period. At best, that explanation is strained. In any event, our results indicate that there is additional important work to be done in this area.

As suggested in Section I, the impact of sellers' concentration on trade restrictions is ambiguous. In his study of the structure of tariffs in Canada, Caves (1976) found a negative and significant relationship between Canadian tariff rates and the four-firm concentration ratio. While the concentration ratio enters the tariff regressions with insignificant and positively signed coefficients, concentration enters the

[7] Two other variables not reported on here: The number of scientists and engineers in research and development relative to total employment (SKLPC) and the wage rate for production workers (WGPRMH) were both insignificant and often positively signed in all of the trade restriction equations.

[8] The index of nontariff barriers to trade is sufficiently subjective in nature and the results obtained in tables 2 and 3 are sufficiently similar that extended discussion of the results in table 2 would not be productive.

nontariff trade barriers regressions with a negative significant coefficient.

Caves speculated that a negative relationship between trade restrictions and seller concentration might reflect a bias by politicians to provide protection to industries with many producers in which the benefits would be widely dispersed. In order to test that hypothesis along with the hypothesis that protection has been sequential with nontariff trade restrictions supplementing predetermined tariffs, we ran the simple simultaneous model summarized in table 4. The results presented in table 4 are obtained from two-stage least-squares regression models. The two models differ only in terms of the designated measure for tariff rates. In the first model, we used the weighted-average tariff measure. In the second model, we used the simple average tariff measure. In both models we used the simple 0-1 dummy measure of nontariff barriers to trade. The results obtained in table 4 are qualitatively invariant with the choice of the subset of comparative advantage variables included. Therefore, for the sake of brevity, the estimators presented in table 4 are offered as representative. The principal findings are that tariffs do positively and significantly affect nontariff trade restrictions, while nontariff trade restrictions have no significant impact on tariff determination, supporting our sequential model, and that nontariff trade restrictions are negatively and significantly related to both seller concentration and the geographical concentration of production, GEOG.[9] The negative sign on GEOG means that nontariff trade restrictions are biased toward industries in which production is distributed across regions of the United States in a manner similar to the distribution of population and therefore toward industries with substantial representation in Congress.

We used a measure of scale economies in U.S. manufacturing in 1967, ESCAL(67), as a proxy for the domestic, ϵ_D, elasticity of supply.

[9] GEOG is measured as follows: index value =

$$\sum_{i=1}^{4}\left|\frac{VS_i}{\sum_{i=1}^{4} VS_i} - \frac{P_{op_i}}{\sum_{i=1}^{4} P_{op_i}}\right|,$$

where VS_i = value of shipments in region i, $i = 1, \ldots, 4$, and P_{op_i} = population of region i, $i = 1, \ldots, 4$. To the extent that production in industry i is distributed across the north, south, central, and western United States in a fashion similar to the general population, the index value will approach zero. The basic data for the index are from the *Census of Manufactures* for 1967. The lack of concordance between TCSIC and SIC data means that 27 observations were lost and that table 4 was estimated using 198 rather than 225 observations. Apart from the fact that the inclusion of GEOG in the regression runs for tables 1–3 would have similarly reduced the sample size from 225 to 198, the negative significance of GEOG in the nontariff dummy equations in table 4 indicates that it should have been included in the earlier estimates as well.

TABLE 4*

SIMULTANEOUS ESTIMATES OF TARIFF AND NONTARIFF TRADE RESTRICTIONS IN THE U.S., 1970

	WEIGHTED-AVERAGE TARIFFS		SIMPLE AVERAGE TARIFFS	
	Dependent Variables			
	U.S. Weighted-Average Tariffs (4.1)	Dummy Variable for Nontariff Barriers (4.2)	U.S. Simple Average Tariffs (4.3)	Dummy Variable for Nontariff Barriers (4.4)
INDEPENDENT VARIABLES	OLSQ	Probit	OLSQ	Probit
Constant	9.69	−2.65	16.02	−7.22
	(2.76)	(1.34)	(4.08)	(2.08)
FTDNMOS	...	$-.18 \times 10^{-2}$	...	$-.23 \times 10^{-3}$
		(1.61)		(1.96)
CONR4L	.03	−.02	.03	−.02
	(1.97)	(2.81)	(1.41)	(2.91)
SKILLD	−23.52	10.80	−27.02	14.55
	(4.49)	(1.97)	(4.61)	(2.15)
SEINRD	...	−0.60	...	0.21
		(0.34)		(0.13)
ESCAL(67)	−4.24	...	−4.83	...
	(1.04)		(1.06)	
LABINT	10.19	−7.42	.79	−.35
	(2.14)	(4.02)	(.15)	(3.41)
KLRA	$-.2 \times 10^{-4}$	$.1 \times 10^{-4}$	$-.2 \times 10^{-4}$	$.13 \times 10^{-4}$
	(1.77)	(1.88)	(1.56)	(2.21)
WDAVPD	...	−.44	...	−.46
		(2.29)		(2.38)
ΔIMP	...	$-.9 \times 10^{-2}$	...	−.29
		(.33)		(1.01)
ΔCON	...	.55	...	.81
		(1.17)		(1.56)
PUSWTTF	...	.52	...	...
		(2.87)		
PUSSMPTAR†	...	...	...	.64
				(2.87)
PDUM2†	.82	...	.27	...
	(.94)		(.28)	
GEOG	1.10	−.92	1.73	−1.53
	(.94)	(2.00)	(1.32)	(2.55)
R^2	.33	...	.27	...
F-statistic	13.51	...	10.25	...
Pseudo-R^2	...	.31	...	.31
Likelihood ratio test	...	49.10	...	49.10

* In converting the geographical dispersion index from SIC to TCSIC codes, 27 of the 225 observations had to be deleted for missing data. Consequently, the models estimated in table 4 are based on 198 observations.

† PUSWTTF, PUSSMPTAR, and PDUM2 are the predicted values of USWTTF, USSMPTAR, and DUM2 on the instrumental variables from the first stage of the two-stage least-squares estimation procedures.

The discussion of Section I suggested that ESCAL(67) should be negatively and significantly related to tariff and nontariff trade restrictions if the government is strongly concerned about the overall welfare costs of protection. ESCAL(67) appears in all of the weighted tariff and nontariff trade barrier dummy regressions with the expected negative sign and generally significant coefficient. While the sign of ESCAL(67) is negative, it is less significant in the simple tariff regressions.

In Section I we indicated that we expected to find a negative and significant coefficient for the absolute value of the own price elasticity of demand for the product in the trade restriction equations. On the assumption that the own price elasticity of demand is reduced by product heterogeneity, we used a measure of product differentiation, WDAVPD, as an inverse measure of the absolute value of the own price elasticity of demand. Consequently, we expected to find a positive and significant coefficient on the product differentiation variable in the tariff and nontariff trade barrier regressions. The relative negative significance of WDAVPD in the nontariff trade barrier regressions compared with the tariff regressions may be partially related to another factor that we have not discussed. Clearly, nonprice, quantitative restrictions are more easily applied and enforced if products are fairly standardized. Consequently, the negative coefficient on product heterogeneity may partially reflect an administrative preference by the government for nonprice restrictions for homogeneous products.

Earlier we argued that while market conditions change continuously, it may still make sense for industry interest groups to push for changes in trade restrictions in discrete fashion. If so, we would expect tariffs to be relatively more productive in industries that are expected to contract and nontariff trade barriers to be relatively more productive in expanding industries. We used the percentage growth in apparent domestic consumption between 1965 and 1970, measured by the percentage growth in domestic shipments plus imports minus exports, as a proxy for demand shifts, ΔD, and the percentage growth in imports between 1965 and 1970 as a proxy for supply shifts, ΔS. Clearly, both the apparent domestic consumption variable and the import expansion variable reflect both demand and supply phenomena and in that sense are poor proxies for the effects of pure demand and supply shifts that we are trying to measure. With that caveat in mind, we simply report that neither of the variables included to measure dynamic influences has any significant explanatory power in the tariff or nontariff trade barrier regressions.

Unfortunately, we have no good measures of the industry-specific revenue effects associated with tariff and nontariff trade restrictions,

or of the administrative costs of implementing alternative forms of trade restrictions across industries. Consequently, all of our estimates suffer from the potential biases associated with important left-out variables.

III. Conclusions

We presented evidence that both tariff and nontariff trade restrictions are found predominantly in industries in which the United States has no comparative advantage vis-à-vis the rest of the world and away from industries in which the deadweight losses to consumers from protection would be high. In addition, we presented direct evidence that nontariff trade restrictions have been used to supplement tariffs and, thereby, offset the trade-liberalizing effects of post–World War II tariff agreements.

We also presented evidence that there are significant differences in the industrial characteristics of industries with tariff protection compared with those with nontariff trade protection. Specifically, tariffs are positively and significantly related to labor intensity and negatively and significantly related to the capital/labor ratio while just the opposite is true for nontariff trade restrictions. In addition, nontariff trade restrictions are negatively and significantly related to both seller concentration and geographical concentration in an industry while both characteristics are positively and insignificantly related to tariffs.

Appendix

Definitions of Variables

Independent Variables

FTDNMOS—Product cycle proxy: unweighted average of Schedule B first trade dates corresponding to TCSIC as of January 1974.

CONR4L —Concentration Ratio, 1970: percentage of shipments accounted for by the four largest firms in the industry.

SKILLD —Skills measure, 1970: professional and kindred workers, plus managers and administrators (except farm), plus craftsmen and kindred workers, as a percentage of total employment. Based on three-digit SIC data with values repeated at four-digit levels.

SEINRD —Percentage of scientists and engineers in R & D in 1970. Based on two-digit SIC data with values repeated at the four-digit level.

ESCAL(67)—Economies of Scale measure, 1970: value of the exponent in the regression equation $V = KN^a$, where V is the ratio of value added in plants employing N persons to average values added for the industry, and K is a constant.

LABINT —Labor Intensity Ratio, 1970 (payroll divided by value added).
KLRA —Capital/Labor Ratio, 1970: Total capital stock divided by employment (thousands of dollars).
WDAVPD —Weighted Average of Product Differentiation measure, 1970: The measure is the coefficient of variation in the unit values of exports destined to different countries weighted by U.S. export shares, i.e., the standard deviation of U.S. export unit values weighted by export shares divided by means of unit values.
ΔIMP —Percentage change in imports between 1965 and 1970.
ΔCON —Percentage change in apparent U.S. consumption between 1965 and 1970, where apparent U.S. consumption is measured by the value of U.S. shipments plus imports minus exports.
GEOG —See n. 9.

Dependent Variables

Tariffs—USWTTF—U.S. weighted-average tariffs for 1970 and USSMPTAR—U.S. simple average tariffs for 1970 are taken from the GATT *Tariff Study*, Geneva, 1970.

Index of Nontariff Barriers—Index of the Incidence of Nontariff Barriers in the U.S., 1970: The index was constructed by the U.S. Tariff Commission and is explained in *Trade Barriers*, Report to the Committee on Finance of the U.S. Senate, Part 2, pp. 160–72, Washington, D.C., April 1974. See n. 6.

Dummy Variable for Nontariff Barriers in the U.S., 1970—DUM2—The value of the dummy variable was set equal to 1 if the index of nontariff barriers was nonzero for a particular product and zero otherwise.

References

Baldwin, Robert E. "Determinants of the Commodity Structure of U.S. Trade." *A.E.R.* 61 (March 1971): 126–46.

———. "Reply." 62 (June 1972): 465.

———. "Trade and Employment Effects in the United States of Multilateral Tariff Reductions." *A.E.R. Papers and Proc.* 66 (May 1976): 142–48. (*a*)

———. "The Political Economy of Postwar U.S. Trade Policy." *Bulletin*, New York Univ. Graduate School of Business, no. 4 (1976). (*b*)

Baldwin, Robert E., and Lage, Gerald M. "A Multilateral Model of Trade-balancing Tariff Concessions." *Rev. Econ. and Statis.* 53 (August 1971): 237–44.

Caves, Richard E. "Economic Models of Political Choice: Canada's Tariff Structure." *Canadian J. Econ.* 9 (May 1976): 278–300.

Caves, Richard E., and Jones, Ronald W. "Tariff Policy and Trade Liberalization." *World Trade and Payments: An Introduction.* 2d ed. Boston: Little, Brown, 1977.

Cheh, John H. "United States Concessions in the Kennedy Round and Short-Run Labor Adjustment Costs." *J. Internat. Econ.* 4 (November 1974): 323–40.

———. "A Note on Tariffs, Nontariff Barriers, and Labor Protection in United States Manufacturing Industries." *J.P.E.* 84, no. 2 (April 1976): 389–94.

Fishelson, Gideon, and Flatters, Frank. "The (Non)Equivalence of Optimal

Tariffs and Quotas under Uncertainty." *J. Internat. Econ.* 5 (November 1975): 385–93.

Helleiner, G. K. "The Political Economy of Canada's Tariff Structure: An Alternative Model." *Canadian J. Econ.* 10 (May 1977): 318–26.

Hufbauer, Gary C. "The Impact of National Characteristics and Technology on the Commodity Composition of Trade in Manufactured Goods." In *The Technology Factor in International Trade*, edited by Raymond Vernon. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1970.

Kreinin, Mordechai E. "More on the Equivalence of Tariffs and Quotas." *Kyklos* 23, fasc. 1 (1970): 75–79.

Pincus, J. J. "Pressure Groups and the Pattern of Tariffs." *J.P.E.* 83, no. 4 (August 1975): 757–78.

Ray, Edward John. "The Optimum Commodity Tariff and Tariff Rates in Developed and Less Developed Countries." *Rev. Econ. and Statis.* 56 (August 1974): 369–77.

Stern, Robert M. "The U.S. Tariff and the Efficiency of the U.S. Economy." *A.E.R. Papers and Proc.* 54 (May 1964): 459–70.

Stone, Joe A. "A Comment on Tariffs, Nontariff Barriers, and Labor Protection in United States Manufacturing Industries." *J.P.E.* 86, no. 5 (October 1978): 959–62.

Inflation, Corporate Income Taxation, and the Demand for Capital Assets

Richard W. Kopcke

Federal Reserve Bank of Boston

The demand for capital is not systematically related to either the level or the rate of change of "effective" income tax rates on corporate capital assets. Rising inflation during the last 10 years has raised the user cost of capital for durable assets relative to that for short-lived assets even though this inflation has raised effective tax rates for more durable capital less than for short-lived assets. Even with replacement-cost depreciation allowances, the level and pattern of investment incentives probably will continue to vary with the inflation rate.

The past decade of high and rising inflation rates has encouraged the study of U.S. income tax codes and inflation's influence on the demand for capital assets. Many have examined the "effective income tax rate" on capital to describe how inflation has altered both the level and pattern of investment incentives. In a recent article, for example, Auerbach (1979) observes: "With a positive rate of inflation, a historic cost depreciation rule biases the choice of asset life toward greater durability" (p. 621). Such a conclusion, based on the behavior of effective income tax rates, contradicts the spirit, if not the letter, of much traditional analysis of investment incentives and the demand for capital (see, e.g., Black 1959; Brown 1962; Samuelson 1964; Hall and Jorgenson 1967; Musgrave and Musgrave 1976; and Boadway 1978) so it warrants examination.

One principal conclusion of this paper is that the demand for capital is not systematically related either to the level or to the rate of

The analysis and conclusions are not necessarily endorsed by the Federal Reserve Bank of Boston or the Federal Reserve System.

[*Journal of Political Economy*, 1981, vol. 89, no. 1]

change of the effective corporate income tax rates on capital assets. Another principal conclusion is that the rising domestic inflation rate of the past 10 years not only has reduced investment incentives generally, but it has reduced the incentive to acquire more durable capital goods most severely. Even if U.S. corporate income tax codes were amended to embrace replacement-cost depreciation allowances, investment incentives would not necessarily be independent of asset service lives nor would the relative demand prices of capital goods necessarily be independent of the inflation rate.

I. Inflation and the Demand for Capital

Assuming perfect markets and certainty of foresight, investors maximize their wealth by purchasing capital yielding no less than ρ, the real discount rate for after-tax corporate cash flow. In other words, the real price of capital, P (the output price is the numeraire),[1] equals

$$\int_0^\infty e^{-\rho t}[(1-\tau)e^{-\delta t}Q_k + \tau PD(t,\delta)e^{-\Pi t}]dt + P\cdot ITC, \text{ or}$$
$$P = \frac{(1-\tau)}{(\rho+\delta)}Q_k + \tau PZ + P\cdot ITC, \tag{1}$$

where τ is the corporate tax rate, δ is the rate of decay of capital, Q_k is the marginal physical product of a unit of new capital, $D(t, \delta)$ is the schedule of depreciation allowances per dollar of investment undertaken t years ago, ITC is the rate of investment tax credit, Z is the present value of depreciation allowances per dollar of investment, and Π is the inflation rate. Formula (1) represents the present value of after-tax operating income plus the present value of the tax shelter afforded by depreciation allowances and tax credits; $D(t, \delta)$ is "discounted" by Π because depreciation allowances are linked to the original purchase price of capital assets, so the real value of these allowances must decline as nominal capital goods prices rise.

In equilibrium, then,

$$CS = P_s(\rho+\delta_s)(1-\tau Z_s)/(1-\tau) = Q_{ks},$$
$$CE = P_e(\rho+\delta_e)(1-\tau Z_e - ITC)/(1-\tau) = Q_{ke}, \tag{2}$$

or the user cost of capital (Hall and Jorgenson 1967)—CS for nonresidential structures, CE for equipment—equals the marginal physical product of capital. Because structures and equipment are each

[1] For the most part, the notation matches that in Auerbach's (1979) article. P may equal unity as it does in Auerbach's article; however, it is often useful to relax this assumption because the relative prices of capital assets are not constant.

represented by a variety of capital goods bearing different rates of decay while offering different flows of capital services, $A(\delta)Q_k$, equilibrium for both equipment and structures demand prevails when

$$\left[\frac{CS(\delta_s)}{A(\delta_s)}\right] \Big/ \left[\frac{CE(\delta_e)}{A(\delta_e)}\right] = Q_{ks}/Q_{ke}. \tag{3}$$

Referring to the equilibrium condition described in expression (2), it is evident that a changing inflation rate, Π, can alter the demand for capital. Higher inflation rates, other things equal, depress the present value of real depreciation allowances; this, in turn, raises the user cost of capital. From (3), it is also evident that if higher inflation rates do not increase the user costs of all capital goods proportionately, the pattern of investment demand will change as well.

In equilibrium,

$$\frac{CS(\delta_1)}{CS(\delta_2)} = \frac{A(\delta_1)}{A(\delta_2)}. \tag{4}$$

Should changing tax rates tend to increase $CS(\delta_1)$ relative to $CS(\delta_2)$, $P_s(\delta_1)$ and $P_s(\delta_2)$ will compensate so that equilibrium is maintained. Of course, if the new balance requires the price of more durable structures, $P_s(\delta_1)$, to decline relative to that of less durable structures, $P_s(\delta_2)$, and if the supply of structures varies with price, then the stock of nonresidential structures will become less durable. The influence of inflation on the demand price of capital and the choice of asset life depends on the sensitivity of the present value of depreciation allowances to changing inflation rates, $\partial Z/\partial \Pi$, as well as the importance of these allowances as a tax shelter, $(1 - \tau Z - ITC)$.

Table 1 shows how rising inflation rates reduce the present value of real depreciation allowances for a variety of capital goods according to current U.S. corporation tax law. The table illustrates two important points. First, the value of depreciation allowances is lower for longer-lived capital assets. Second, with higher inflation, the value of depreciation allowances drops at a faster rate for more durable capital. Therefore, the modest role of depreciation allowances in the cash flow accruing to durable capital does not mean that these assets are necessarily less sensitive to changing inflation rates than are less durable assets.

In fact, as shown in table 2, the tax shelter offered by depreciation allowances is sufficiently important for long-lived capital that higher inflation rates increase the user cost of capital for these assets more than that of less durable assets. As the annual inflation rate rises from zero to 6 percent, other things equal, the investment incentives for 40-year structures fall 22 percent, but the incentives for 20-year structures fall only 19 percent. Investment incentives generally drop much less for the tabulated categories of equipment.

TABLE 1

THE PRESENT VALUE OF REAL DEPRECIATION ALLOWANCES PER DOLLAR OF INVESTMENT

A. PRODUCERS' DURABLE EQUIPMENT*

Inflation Rate (%)	Asset Lifetime (Years)		
	5	10	15
0	.44	.43	.42
2	.42	.40	.38
4	.41	.37	.34
6	.39	.35	.31
8	.37	.33	.29
10	.36	.31	.27
20	.30	.23	.19
30	.25	.18	.14

B. NONRESIDENTIAL STRUCTURES†

Inflation Rate (%)	Asset Lifetime (Years)		
	20	30	40
0	.41	.38	.36
2	.35	.31	.28
4	.30	.26	.22
6	.27	.22	.19
8	.24	.19	.16
10	.22	.17	.14
20	.15	.10	.09
30	.11	.08	.06

NOTE.—Entries equal nominal depreciation allowances multiplied by .46, the marginal corporate income tax rate, discounted by $(1.015)(1 + \Pi)$, where 1.5 percent is the real after-tax risk-adjusted discount rate (see Brown 1962) and Π is the inflation rate.

* Nominal depreciation allowances are calculated using the sum-of-the-years'-digits formula.

† Nominal depreciation allowances are calculated using the 150 percent–declining formula with a switch to straight-line allowances after one-third of the asset's life has passed.

Because rising inflation reduces the present value of depreciation allowances most quickly for the longest-lived assets, however, the user cost of capital for these assets becomes progressively less sensitive to further increases in the inflation rate. For example, if the inflation rate were to rise from 6 to 10 percent, investment incentives would drop another 5 percent for 40-year structures compared with an incremental decline of 6 percent for 20-year structures. Should the inflation rate exceed 20 percent, the erosion of real depreciation allowances would be so severe for the shorter-lived structures that longer-lived structures become relatively more attractive investments.

Despite a popular contention that the value of the depreciation tax shelter, hence its annual loss in value due to inflation, is greatest for

TABLE 2

INFLATION AND INVESTMENT INCENTIVES

A. PRODUCERS' DURABLE EQUIPMENT

INFLATION RATE (%)	ASSET LIFETIME (Years)		
	5	10	15
0	1.00	1.00	1.00
2	.96	.94	.92
4	.93	.89	.86
6	.90	.85	.81
8	.87	.81	.78
10	.85	.78	.75
20	.77	.70	.67
30	.72	.65	.63

B. NONRESIDENTIAL STRUCTURES

INFLATION RATE (%)	ASSET LIFETIME (Years)		
	20	30	40
0	1.00	1.00	1.00
2	.91	.89	.88
4	.85	.83	.82
6	.81	.79	.78
8	.78	.76	.76
10	.76	.74	.74
20	.69	.69	.70
30	.67	.67	.68

NOTE.—Entries for each asset equal the reciprocal of the user costs of capital, expressions (2), divided by the reciprocal of user costs when the inflation rate equals zero. Values of Z_e and Z_s are taken from table 1, and, for equipment, ITC equals .1 except for 5-year assets for which ITC equals .067. The figures in this table are independent of the pattern of "true" capital decay—inflation alters the relative user cost through the factor $(1 - \tau Z - ITC)$.

less durable assets, it is the percentage change in the effective purchase price of capital, $P(1 - \tau Z - ITC)$, that influences investment incentives. Table 2 shows that higher inflation rates raise this effective purchase price most for more durable, not less durable, assets for relevant inflation rates in the United States. The more durable the asset, the more its depreciation allowances are postponed, hence the more inflation may "discount" the value of these allowances before they may be claimed, as shown in the first table. Consequently, high inflation rates discourage the purchase of longer-lived capital more than short-lived assets.[2] Ironically, the class of investment goods most

[2] This is true whenever the pattern of investment demand for high inflation rates is compared with the pattern of demand for inflation rates near zero. As explained in the text, however, as the inflation rate rises from an already high level, the relative demand price of some longer-lived assets may rise.

needing "protection" from the high inflation rates of the past 15 years—nonresidential structures—has not received the benefit of the tax credits and accelerated depreciation granted to producers' durable equipment.

This shift in the relative capital costs that inflation induces will alter investment demand to favor less durable capital goods, thereby depressing the relative demand price for more durable capital assets. As a result, the stock of more durable structures, other things equal, will fall relative to that of less durable structures. To the extent that structures and equipment are substitutes, the stock of structures generally will decline relative to that of equipment. Of course, the higher user cost for all capital goods implies a declining capital-labor ratio in addition to this substitution among varieties of capital assets.

Assuming that the real price for all capital goods equals unity, a unit of output may be consumed or installed either as short-lived equipment or as a very durable structure; consequently, changing rates of inflation cannot affect asset prices. Rather, from expression (4), if $CS(\delta)/A(\delta)$ for a more durable factory exceeds that for a less durable factory due to increasing inflation, investors will install less durable structures; more durable factories are now economically obsolete.

In summary, with current U.S. income tax codes, which link depreciation allowances to the original purchase price of capital assets, changing inflation rates alter the relative user costs of capital, other things equal. Higher inflation rates initially raise these user costs more for long-lived assets than for short-lived assets, thereby reducing the average service life of the capital stock.

II. Inflation and the Effective Corporate Tax Rate

Auerbach (1979) defines the investor's implicit discount rate as

$$v \equiv c - \delta, \tag{5}$$

and then the effective corporate tax rate,

$$\theta \equiv (v - \rho)/v. \tag{6}$$

In view of expressions (2) and (3), it is possible for θ either to rise or to fall with the service lives of capital assets. Consider, however, the simple case wherein $A(\delta)$ equals unity and prices, $P(\delta)$, are flexible for all assets (the definition of a unit of capital has eliminated the apparent differences in productivity among pure substitutes); then (from [4]), $CS(\delta_1) = CS(\delta_2)$ and $\theta(\delta_1) > \theta(\delta_2)$: The effective tax rate on the more durable factory is largest, though the user costs are equal.

As explained in the previous section, a higher inflation rate will tend to raise the user cost for long-lived structures more than that of less durable structures. A higher inflation rate, then, will also increase

Q_{ks} by reducing the stock of nonresidential structures, other things equal. Retaining the assumption that $A(\delta)$ equals unity, from (4) $CS(\delta_1)$ will still equal $CS(\delta_2)$ in equilibrium, and both will be greater, with higher inflation rates as long as capital goods prices are flexible. Though $CS(\delta_1)$ and $CS(\delta_2)$ rise proportionately, $\theta(\delta_1)$ increases less than $\theta(\delta_2)$ because δ_1, δ_2, and ρ do not change. Even though the effective tax rate, θ, on less durable capital rises more than that on longer-lived assets, the capital stock is comprised of more short-lived assets with higher inflation rates because $P_s(\delta_1)$ (the price of the more durable asset) has fallen relative to $P_s(\delta_2)$ in equilibrium, as described after expression (4).

If $A(\delta)$ rises with δ and if the prices of all capital were unity, a medium-lived asset may have a relatively high user cost and yet be the asset in greatest demand due to its high productivity, $A(\delta)$. Consequently, the effective tax rate, θ, for this asset may exceed that for many alternative assets not favored by investors.

Assume, for $\delta_1 < \delta_2 < \delta_3$, that

$$\frac{CS(\delta_1)}{A(\delta_1)} > \frac{CS(\delta_2)}{A(\delta_2)} < \frac{CS(\delta_3)}{A(\delta_3)}, \text{ and} \tag{7}$$

$$CS(\delta_1) < CS(\delta_2) < CS(\delta_3).^3$$

It is reasonable to assume $CS'(\delta) < 1$ (see n. 3); therefore,

$$\begin{aligned}\theta(\delta_1) &= [CS(\delta_1) - \delta_1 - \rho]/[CS(\delta_1) - \delta_1] > \\ \theta(\delta_2) &= [CS(\delta_2) - \delta_2 - \rho]/[CS(\delta_2) - \delta_2] > \\ \theta(\delta_3) &= [CS(\delta_3) - \delta_3 - \rho]/[CS(\delta_3) - \delta_3].\end{aligned} \tag{8}$$

The medium-lived asset is the optimal investment according to (7), but expression (8) reveals that its effective tax rate is not the lowest.

Just as the level of θ is a poor guide to investment incentives, so the change in the effective tax rate is a poor guide to the change in investment incentives. From (7), the short-lived asset (δ_3) will be preferred to the medium-lived asset once inflation is high enough to cause

$$\frac{CS(\delta_3)}{CS(\delta_2)} < \frac{A(\delta_3)}{A(\delta_2)}, \tag{9}$$

because $\partial[CS(\delta_3)/CS(\delta_2)]/\partial\Pi < 0$.[4]

[3] $CS'(\delta) = [(1 - \tau Z) - \tau(\rho + \delta)Z'(\delta)]/(1 - \tau)$. According to table 1, $\tau Z'(\delta) \approx 2$, so the first term in the brackets is approximately 0.6 and the second is approximately 0.3; therefore, it is reasonable to assume $0 < CS'(\delta) < 1$. Similarly, $CS''(\delta) < 0$.

[4] As shown in table 2, unless inflation rates exceed 20 percent, a higher inflation rate raises the user cost for less durable structures (δ_3) proportionately less than the user cost of more durable structures (δ_2). The relative user cost of less durable structures declines as inflation rises.

Though rising inflation depresses the user cost of short-lived assets relative to that of longer-lived assets, however, it need not depress the relative effective tax rate on shorter-lived assets: $\partial\theta(\delta_3)/\partial\Pi$ is not necessarily less than $\partial\theta(\delta_2)/\partial\Pi$, nor is $\partial[\theta(\delta_3)/\theta(\delta_2)]/\partial\Pi < 0$ always. From expressions (7) and (8), for example,

$$\left[\frac{\partial\theta(\delta_3)}{\partial\Pi}\right] \bigg/ \left[\frac{\partial\theta(\delta_2)}{\partial\Pi}\right] = \frac{\partial \ln [CS(\delta_3)]/\partial\Pi}{\partial \ln [CS(\delta_2)]/\partial\Pi} \cdot \frac{CS(\delta_3)}{CS(\delta_2)} \cdot \frac{[CS(\delta_2) - \delta_2]^2}{[CS(\delta_3) - \delta_3]^2}. \quad (10)$$

The product of the last two factors in the right-hand side of (10) exceeds unity;[5] consequently, the left-hand side of (10) may exceed unity—the tax rate on less durable assets may rise more than that on longer-lived assets—even though $CS(\delta_3)/CS(\delta_2)$ is falling. Whenever $\theta(\delta_3)$ rises more than $\theta(\delta_2)$, $\theta(\delta_3)/\theta(\delta_2)$ also increases because $\theta(\delta_3) < \theta(\delta_2)$ in this example. While relative user costs for shorter-lived structures fall with rising inflation, the effective tax rate on these assets may rise more than that on longer-lived structures.

Finally, from (2), (5), and (6), assuming $P \equiv 1$,

$$\theta = [\tau(1 - Z) - ITC]/[1 - \tau Z - ITC - (1 - \tau)\delta/(\rho + \delta)]. \quad (11)$$

Assuming that the intertemporal discount rate (ρ) is constant, the effective tax rate on corporate capital must vary with the inflation rate unless the statutory tax rate (τ) equals zero or depreciation allowances are adjusted for changing capital goods prices—$D(t, \delta)$ is replaced by $D(t, \delta)e^{\Pi t}$ in (1)—so that $\partial Z/\partial\Pi = 0$; $D(t, \delta)$ may correspond to any capital consumption scheme, including "economic depreciation."

It is likely, however, that ρ changes whenever the inflation rate changes;[6] therefore, the demand for capital may change whenever Π

[5] See n. 3. Because $CS'(\delta) > 0$, $CS(\delta_3) > CS(\delta_2)$. $CS'(\delta) < 1$, however, implies that $[CS(\delta_2) - \delta_2]^2 > [CS(\delta_3) - \delta_3]^2$. Therefore, both of the last two factors of (10) exceed unity.

[6] In many standard neoclassical growth models, combining intertemporal consumption and production decisions (e.g., Sidrauski 1967; Intriligator 1971), ρ by hypothesis is constant. In other versions of this model the intertemporal utility function is replaced by a saving rate and portfolio balance relationship (Tobin 1955, 1965). In these latter models, higher inflation rates can lower the real return on money balances, thereby encouraging investors to acquire more capital, by reducing ρ, until its marginal physical product drops enough to restore portfolio equilibrium. In other models, ρ is neither constant nor essentially equal to the real return on money. Inflation is a by-product of public and private policies intended to redistribute resources (Schumpeter 1939; Robinson 1954, 1962; Pasinetti 1962). This reallocation of wealth, income, or spending undoubtedly will change the social rate of time preference, ρ. Although Tobin's growth model suggests ρ would fall as inflation increases, Schumpeter's model suggests ρ would rise. It is not unreasonable, then, to assume that higher inflation depresses ρ and, in turn, increases the average service life of capital if corporate income taxes are indexed. Yet, so little is known about the social rate of time preference that it is unreasonable to ignore the alternative conclusion: Higher inflation rates accompany an increase in ρ, thereby depressing the average service life of capital.

changes, unless this demand is also independent of ρ. By setting $ITC = 0$ and either setting $\tau = 0$ or allowing depreciation schedules that are a linear combination of "economic depreciation" adjusted for changing capital goods prices and immediate expensing of investment expenditures, not only is the effective tax rate independent of the inflation rate, but it is independent of the discount rate, ρ, and asset service lives. (Even if depreciation allowances were "indexed" for inflation, unless these allowances correspond to the pattern of an asset's "economic depreciation," the effective tax rate will not be independent of the asset's service life [see Samuelson 1964].) Of course, these last reforms accomplish this goal by pegging the effective tax rate at zero. Nevertheless, even though $\theta = 0$ for all values of Π, δ, and ρ, these reforms cannot make the user cost of capital independent of ρ and δ because (from [2]):

$$C = (\rho + \delta). \tag{12}$$

The demand for capital, therefore, is not systematically related either to the level or to the rate of change of the effective corporate tax rate defined by expressions (5) and (6). The demand for capital depends on the user cost of capital and the marginal physical product of capital alone in neoclassical macro models, so the income tax codes influence investment incentives only by altering capital's user cost. For studying the demand for capital, then, the suggested effective corporate tax rate, θ, is not a sufficient statistic—it alone is not a definitive measure of investment incentives. Conversely, the user cost of capital does not measure the effective tax burdens borne by the various capital assets. Each concept—the cost of capital and the effective tax rate—serves a particular purpose, and their roles are distinct.

III. Summary

Unless the aggregate rate of time preference has declined substantially, the rising inflation rate of the past 10 years not only has reduced business's demand price for capital generally but it has reduced the relative demand price for more durable assets most severely. Even if the corporate tax rate were zero or depreciation allowances were a linear combination of economic depreciation adjusted for changing capital goods prices and immediate expensing of investment expenditures, thereby pegging the effective corporate income tax rate at zero, the level and pattern of investment incentives most likely would vary with the inflation rate.

Of course, it is one thing for inflation to influence investment incentives directly; it is quite another for changing investment incentives to accompany a changing inflation rate because the intertem-

poral discount rate varies with the inflation rate. Although it is probably unreasonable to expect any tax reform to insulate the demand price of capital from the influence of policies that reallocate resources, any schedule of depreciation allowances adjusted for changing prices can insulate investment incentives from inflation per se.

References

Auerbach, Alan J. "Inflation and the Choice of Asset Life." *J.P.E.* 87, no. 3 (June 1979): 621–38.

Black, John. "Investment Allowances, Initial Allowances and Cheap Loans as Means of Encouraging Investment." *Rev. Econ. Studies* 27 (October 1959): 44–49.

Boadway, Robin W. "Investment Incentives, Corporate Taxation and Efficiency in the Allocation of Capital." *Econ. J.* 88 (September 1978): 470–81.

Brown, Edgar C. "Tax Incentives for Investment." *A.E.R. Papers and Proc.* 52 (May 1962): 335–45.

Hall, Robert E., and Jorgenson, Dale W. "Tax Policy and Investment Behavior." *A.E.R.* 57 (June 1967): 391–414.

Intriligator, Michael D. *Mathematical Optimization and Economic Theory.* Englewood Cliffs, N.J.: Prentice-Hall, 1971.

Musgrave, Richard A., and Musgrave, Peggy B. *Public Finance in Theory and Practice.* 2d ed. New York: McGraw-Hill, 1976.

Pasinetti, Luigi L. "Rate of Profit and Income Distribution in Relation to the Rate of Economic Growth." *Rev. Econ. Studies* 29 (October 1962): 267–79.

Robinson, Joan. "The Production Function and the Theory of Capital." *Rev. Econ. Studies* 21, no. 2 (1954): 81–106.

———. *Essays in the Theory of Economic Growth.* New York: St. Martin's, 1962.

Samuelson, Paul A. "Tax Deductibility of Economic Depreciation to Insure Invariant Valuations." *J.P.E.* 72, no. 6 (December 1964): 604–6.

Schumpeter, Joseph A. *Business Cycles.* New York: McGraw-Hill, 1939.

Sidrauski, Miguel. "Rational Choice and Patterns of Growth in a Monetary Economy." *A.E.R. Papers and Proc.* 57 (May 1967): 534–44.

Tobin, James. "A Dynamic Aggregative Model." *J.P.E.* 63, no. 2 (April 1955): 103–15.

———. "Money and Economic Growth." *Econometrica* 33 (October 1965): 671–84.

A Monetary Approach to the Crawling-Peg System: Theory and Evidence

Mario I. Blejer

Hebrew University of Jerusalem

Leonardo Leiderman

Tel-Aviv University

This paper develops and estimates a model of the joint determination of the exchange rate, international reserves, and the rate of inflation under a crawling-peg system. The framework presented, which is an extension of previous work on the monetary approach, generates short-run deviations from purchasing power parity that occur simultaneously with movements in both international reserves and the exchange rate. The model is estimated by full-information maximum likelihood on the basis of quarterly data for Brazil.

With the advent of the modern literature on the monetary approach to the balance of payments and to the exchange rate, and following the seminal contributions of Mundell (1968, 1971), Johnson (1972), and Dornbusch (1973), a large amount of effort has been devoted to emphasizing the role of the money market and asset markets in the determination of the balance of payments and the exchange rate. The

We are also research associates at the Center for Latin American Development Studies of Boston University, whose support in the preparation of this paper is gratefully acknowledged. We are indebted to D. Cavallo, R. Dornbusch, J. A. Frenkel, G. Hanoch, A. Harberger, R. Hodrick, A. Stockman, and referees of this *Journal* for valuable comments. Previous versions of this paper were presented at the P. Sapir Conference on Inflation and Development, Tel Aviv; the Summer Meetings of the Econometric Society, Montreal; and a seminar at the Fundación Mediterranea, Córdoba, Argentina.

[*Journal of Political Economy*, 1981, vol. 89, no. 1]

long-run effects of monetary shocks in a regime of fixed exchange rates, the effects of devaluations and changes in commercial policy in a monetary setting, and the short-run consequences of monetary imbalance on the inflation rate and the balance of payments have recently been the object of careful theoretical and empirical study. The importance of monetary variables in the determination of the exchange rate in a regime of floating rates has also been analyzed in detail.[1]

Most of these studies, however, have dealt either with cases in which the exchange rate has been kept fixed for very long periods or with free-floating cases in which the exchange rate is predominantly determined by the interaction of market forces without government intervention.[2]

Yet in recent years many countries have experienced simultaneous fluctuations in their exchange rates and international reserves. These fluctuations have occurred either under a managed-float system, characterized by government intervention in the foreign-exchange market, or under a crawling-peg system, in which the authorities determine, and periodically change, the country's exchange rate.[3] In order to analyze the experience of countries with these characteristics, an extension to the previous formulations of the monetary approach is required.

In this paper we develop and estimate a model for the analysis of the joint determination of the exchange rate, international reserves, and the rate of inflation under a crawling-peg system. Section I

[1] A number of important contributions on the topic are collected in Frenkel and Johnson (1976, 1978) and in the International Monetary Fund (1977) volume. A review of the empirical evidence on the monetary approach to the balance of payments and exchange rates is presented in Magee (1976). Johnson (1977) and Whitman (1975) present comprehensive reviews of the recent contributions to the monetary-approach literature, while a critique of some of the central aspects of that approach is found in Hahn (1977). All these deal with fixed-exchange-rate regimes. Among those dealing with floating rates are Kouri (1975), Dornbusch (1976), Frenkel (1976), Mussa (1976), and Frenkel and Clements (in press).

[2] Exceptions are the studies by Artus (1976) and by Girton and Roper (1977). The latter develop and test a monetary model of exchange-market pressure, which is defined as a composite variable that includes changes in both the exchange rate and international reserves. This model was recently applied to Brazil by Connolly and da Silveira (1979) who, using annual data, attempt to explain the behavior of the exchange-market pressure variable. As will be seen, our model is designed to deal simultaneously with the joint determination of each of the two components of exchange-market pressure as well as of the domestic rate of inflation.

[3] A theoretical analysis which concludes that the optimum exchange-rate regime will correspond to neither of the extremes of a completely fixed or a completely flexible rate is presented by Boyer (1978) and Frenkel (1980), who concludes that "for the purpose of empirical work it is useful to design a framework of the adjustment mechanism that can accommodate simultaneously changes in international reserves and changes in the exchange rate."

presents the basic model and uses it to study the effect of changes in domestic monetary policy on the evolution of these three endogenous variables. The empirical implementation of the model, presented in Section II, uses quarterly data from Brazil covering the crawling-peg period, 1968–77. The parameters of the model are estimated simultaneously using a full-information maximum-likelihood method. Concluding remarks are presented in Section III.

I. The Model

Monetary Equilibrium

The model is a variant of the monetary approach to the balance of payments. Its main feature is that it accounts for different degrees of exchange-rate flexibility (from a freely floating parity to a completely fixed exchange rate). We take the case of a small country, defined as one whose international price of traded goods is exogenously determined, and we allow for the existence of "nontraded" goods, defined as goods whose price responds, at least in the short run, to domestic monetary variables.

The basic relationships of the monetary sector are

$$M_s \equiv a(R + D), \tag{1}$$

$$M_d = Pm_d, \tag{2}$$

$$m_d = f(y, \pi^e), \tag{3}$$

where M_s is the nominal supply of domestic money; a is the money multiplier; R is the foreign-exchange reserves held by the central bank; D is the domestic-credit component of the monetary base; M_d is the demand for nominal cash balances; P stands for a price index that includes traded and nontraded goods; and m_d is the real demand for money, assumed to be a function of real income,[4] y, and of the alternative cost of holding money, proxied by π^e, the expected rate of inflation.[5]

Although it is possible to postulate a mechanism of lagged adjustment, as is done in the empirical section, here we assume for

[4] We assume that the rate of growth of real income is exogenous with respect to monetary variables (we therefore assume a version of the natural rate hypothesis). Although restrictive, this assumption allows us to focus on the specific effects of money in an open economy in the context of the monetary approach. At the empirical level we found evidence supporting our assumption for the case under consideration. A complete report of the relevant causality tests is available from us on request.

[5] As it happens, the results below are valid under a large variety of assumptions about the mechanism for the formation of π^e, provided that the latter depends on lagged variables. In the empirical applications, we assume that π^e is formed by a version of rational expectations; see Sec. II.

simplicity that the money market clears in each period so that the nominal stock of money is equalized, ex post, with the demand for nominal balances. The assumption requires the existence of the flow equilibrium

$$M_s^* = M_d^*, \tag{4}$$

where the asterisk indicates the percentage rate of change of the variable. Differentiating equations (1) and (2) logarithmically, the flow-equilibrium condition can be rewritten as

$$a^* + (1 - \gamma)R^* + \gamma D^* = P^* + m_d^*, \tag{5}$$

where γ is a factor of proportionality equal to $D/(R + D)$ and P^* is the domestic rate of inflation.

The Domestic Rate of Inflation and the Balance of Payments

When traded and nontraded goods are both present, the domestic rate of inflation can be measured as a weighted average of the rate of change of the price of both kinds of goods,

$$P^* = \lambda P_T^* + (1 - \lambda)P_{NT}^*, \tag{6}$$

where P_T, P_{NT} is the price in domestic currency of traded, nontraded goods, respectively, and λ is the share of traded goods in total expenditure. In a small economy P_T^* is determined by the world rate of inflation (P_w^*) and by the variations in the effective exchange rate (ρ^*):

$$P_T^* = P_w^* + \rho^*. \tag{7}$$

The price of nontraded goods, however, may be affected by domestic factors, at least in the short run. Since an ex ante excess supply of money implies excess demand in the goods market, and if excess demand for nontraded goods varies monotonically with excess demand throughout the economy, we can expect the price of nontraded goods relative to that of traded goods to be a function of imbalance in the money market,

$$P_{NT}^* = P_T^* + \theta\Omega, \tag{8}$$

where Ω is a measure of monetary imbalance and θ is the elasticity of the relative price with respect to the monetary variable.[6]

An important feature of equation (8) is that it includes only con-

[6] The elasticity θ is a function of the elasticity of substitution between traded and nontraded goods in consumption and production and of the income elasticity of the nontraded goods. For a detailed description of the dynamics of domestic-price determination in a monetary model with traded and nontraded goods, see, e.g., Parkin (1974) and Blejer (1977).

temporaneous variables. In practice, however, it may well be that P^*_{NT} depends on lagged P^*_T and Ω reflecting persistence effects. To allow for these effects, it is possible to postulate a relationship of the form

$$(P^*_{NT} - P^*_T)_t = \theta\Omega_t + \alpha(P^*_{NT} - P^*_T)_{t-1}, \tag{9}$$

where t is a time subscript and α is the adjustment coefficient. Empirically, however, we found that our results were quite insensitive to the particular specification adopted. We therefore considered equation (8) as the relevant representation for relative-price determination. The structure of the model resulting from equation (9) and the results obtained using that specification are presented in the Appendix.[7]

As regards the measure Ω, it is important to remember that a central conclusion of the monetary approach to the balance of payments is that in a small open economy the nominal supply of money may be beyond the control of the monetary authority. Under fixed exchange rates the monetary authority can only determine the ex ante quantity of money by changing the domestic-credit component of the base or by manipulating the money multiplier. In conjunction with the flow demand for real balances generated by adjustments in the desired stock, such measures create an ex ante excess flow supply of money to which the public reacts by changing the level of the international reserve component of the base through the balance of payments and by affecting the rate of domestic inflation. The ex post nominal quantity of money in an open economy is then influenced by the public's response to ex ante conditions in the money market.[8]

It appears, therefore, that the relevant measure to account for the monetary effects on the goods market in an open economy should be an ex ante measure which does not include the endogenous reaction of the foreign component of the base. For that reason we define Ω in equation (8) as the gap (in percentage terms) between the ex ante change in the money supply (i.e., a change in the domestic-credit component of the base and in the money multiplier)[9] and changes in demand. Equation (8) can therefore be rewritten as

$$P^*_{NT} = P^*_T + \theta(\gamma D^* + a^* - M^*_d). \tag{10}$$

[7] Eq. (8) ignores possible real sector effects on the relative price of nontraded goods (e.g., real wages, income policies). Although the equation can be modified to take these factors into account, this would result in a framework far beyond our intended extension of the monetary approach to the case of a crawling peg.

[8] For time-series analysis of European data and econometric tests of the interaction between changes in domestic credit and in international reserves implied by the monetary approach, see Blejer (1979) and Leiderman (1980).

[9] Empirical examination of the assumption of exogeneity of $(\gamma D^* + a^*)$, on the basis of causality tests, indicates nonrejection of the assumption; complete test results are available from us on request.

Substituting (7) into (10) and then (7) and (10) into (6), we obtain, after some manipulation, the following expression for the rate of domestic inflation:[10]

$$P^* = \epsilon(P_w^* + \rho^*) + (1 - \epsilon)(\gamma D^* - m_d^*), \tag{11}$$

where $\epsilon = [1 + \theta(1 - \lambda)]^{-1}$.

In addition to changes in the price level, there are also changes in international reserves operating to restore monetary equilibrium. An expression for the money account of the balance of payments, which is equal to the change in the international reserves held by the central bank, may therefore be obtained by substituting P^* in equation (5)—the flow equilibrium condition for the money market—for its value in (11) and rearranging the terms:

$$(1 - \gamma)R^* = \epsilon(P_w^* + \rho^* + m_d^* - \gamma D^*). \tag{12}$$

Equations (11)–(12) present the domestic rate of inflation and the balance of payments as functions of world inflation, exchange-rate policy, and the rate of ex ante excess flow supply of money. When nontraded goods are absent ($\lambda = 1$) or when their price is not sensitive to monetary imbalance ($\theta = 0$), then $\epsilon = 1$, and the model is similar to the classical long-run formulation of the monetary approach (see Johnson 1972). In such a case, domestic monetary variables do not affect the domestic rate of inflation which, if the exchange rate is not altered, is fully determined by the world rate, and every ex ante monetary shock will lead to reserve depletion due to a balance-of-payments deficit.

The Endogeneity of the Exchange Rate in a Crawling-Peg System

Except in a fully flexible exchange-rate system (or in a managed float), the exchange rate is regarded by governments as a policy instrument, and its fluctuations are generally influenced by policy decisions aimed at one or more goals. Unlike under an adjustable-peg regime, under a crawling-peg system the exchange rate is changed frequently according to some set of rules adopted in order to attain the government's objectives.[11] The variation of the exchange rate can therefore be considered as following a sort of reaction function which reflects policy goals as well as the parameters of the model adopted and the

[10] Assuming, for simplicity, a constant money multiplier, i.e., $a^* = 0$. In the empirical section, however, changes in the money multiplier are explicitly considered.

[11] A number of proposals for the operation of the crawling peg as well as analyses of the stability of the system have recently been presented in the literature; see, e.g., Williamson (1966), Cooper (1970), Kenen (1975), Levin (1975, 1977), and Mathieson (1976).

values of the exogenous and endogenous variables considered relevant for the desired goals. The goals, and therefore the crawling rules, may vary from country to country.[12]

With a view to the empirical implementation of the model for Brazil, we postulate here that the policy objective is to avoid long-run changes in the real exchange rate and that the nominal rate (ρ) is therefore altered to maintain purchasing power parity.[13] We assume, in addition, that the full adjustment of the exchange rate may take more than one period, and we shall analyze the effects of differences in the speed of exchange-rate adjustment. The reaction function implied by the policy rule is

$$\rho_t^* = \beta \sum_{i=0}^{n} (1 - \beta)^i L^i (P^* - P_w^*)_t, \tag{13}$$

where t is a time subscript, β indicates the portion of the current differential rate of inflation transmitted to the exchange rate in the current period, and L is the lag operator ($L^i x_t = x_{t-i}$).[14]

This formulation specifically assumes that, in addition to the current-period adjustment, the exchange rate will continue to change in each subsequent period by a fraction β of the still unadjusted differential until the whole differential has been transmitted to the exchange rate. Obviously, the greater is β, the faster will the rate adjust. If $\beta = 1$ our model does not differ conceptually from the monetary-approach model of free-floating exchange rates, since the rate of depreciation is then fully determined by the domestic-foreign inflation differential (see Frenkel 1976).

The Functioning of the Model

We proceed now to solve the model for the three endogenous variables in which we are interested: the rate of inflation, the rate of

[12] Kenen (1975) analyzes in detail the relative efficiency of a number of alternative sliding-parity rules. Mathieson (1976) studies the consequences of using a welfare—instead of a balance-of-payments—objective as the guideline for the crawl.

[13] The appropriateness of this assumption for Brazil is discussed in Sec. II. In a previous version of this paper a number of alternative rules were incorporated into the model, among them maintaining a given level of nominal reserves ($R^* = 0$) and maintaining a given level of real reserves ($R^* - P_w^* = 0$). Although the dynamics of the system change with the policy rule, its basic structure is not affected. For presentational convenience the functioning of the model is here confined to a single rule.

[14] A reaction function of this type follows Dean (1974) in the sense that endogenous target variables are a function of the current values of other endogenous variables; this is justified on the grounds that the latter are a plausible representation of the structure of the model and may therefore provide prior knowledge of the structural-form coefficients. This differs from the approach of earlier works such as Friedlaender (1973) where endogenous variables depend only on exogenous or lagged-endogenous variables.

change of foreign-exchange reserves, and the rate of change of the exchange rate. Combine (11) and (13) to solve for ρ_t^*,

$$\rho_t^* = \frac{\beta(1 - \epsilon)}{1 - \beta\epsilon - (1 - \beta)L} (\gamma D^* - m_d^* - P_w^*)_t, \tag{14}$$

and substitute this in (11) and (12) to obtain

$$P_t^* = \frac{\epsilon(1 - \beta)(1 - L)}{1 - \beta\epsilon - (1 - \beta)L} (P_w^*)_t + \frac{(1 - \epsilon)[1 - (1 - \beta)L]}{1 - \beta\epsilon - (1 - \beta)L} (\gamma D^* - m_d^*)_t, \tag{15}$$

and

$$(1 - \gamma)R_t^* = \frac{\epsilon(1 - \beta)(1 - L)}{1 - \beta\epsilon - (1 - \beta)L} (P_w^* + m_d^* - \gamma D^*)_t. \tag{16}$$

The three endogenous variables are functions of current and lagged values of foreign inflation and domestic monetary variables. We can consider now which type of monetary policy will equalize the domestic rate of inflation to the world rate. Because the coefficients of P_w^* and $(\gamma D^* - m_d^*)$ in equation (15) add to unity, P^* will equal P_w^* when the monetary authority expands the money supply at the rate necessary to satisfy the growth in real demand and to replace the depreciated value of the nominal stock. This is achieved when the exogenous component of the supply of money (domestic credit) expands at a rate that exceeds the growth in the demand for real balances by the world rate of inflation, that is, when the ex ante excess flow supply of money is equal to the world inflation rate,

$$(\gamma D^* - m_d^*) = P_w^*, \tag{17}$$

which implies

$$P^* = P_w^* \tag{18}$$

and

$$\rho^* = R^* = 0. \tag{19}$$

To illustrate the functioning of the model further we can analyze the effects of a domestic-credit shock. To do so, it is convenient to consider an economy whose initial equilibrium satisfies equations (18)–(19). Starting from such an equilibrium position, and as long as β and ϵ are smaller than unity, an acceleration in the rate of expansion of the domestic-credit component of the monetary base, $\Delta(\gamma D^*)$, will cause the rate of domestic inflation to depart from the world rate (here assumed constant), the exchange rate will rise, and a balance-of-payments deficit will be created.

If the rate of growth of domestic credit is sustained at the new, higher level, the exchange rate will eventually adjust to account fully for the differential in the inflation rates, which will then be equal to $\Delta(\gamma D^*)$.[15] Once the changes in the exchange rate fully reflect the differential rate of inflation, the excess supply of money created by the government in each period is fully eliminated through an increase in domestic prices, and no further flows of foreign exchange reserves occur. This process is illustrated in figure 1 for different values of β and ϵ.[16] As β increases, for given values of ϵ, the exchange rate adjusts faster to the monetary shock, and domestic inflation both diverges faster from world inflation and converges faster to its new equilibrium rate. As can be observed by comparing paths III and II in figure 1, the balance of payments deteriorates less and returns to equilibrium more rapidly (implying a smaller total loss of reserves) as β increases. When $\beta \rightarrow 1$, the system approaches a flexible-exchange-rate model, like the one presented by Frenkel (1976), in which domestic inflation is always independent of world inflation and the balance of payments is always zero. Letting $\beta = 1$ in equations (14)–(16), we obtain

$$P_t^* = (\gamma D^* - m_d^*)_t, \tag{20}$$

$$\rho_t^* = (\gamma D^* - m_d^* - P_w^*)_t = (P^* - P_w^*)_t, \tag{21}$$

$$R_t^* = 0. \tag{22}$$

The speed of adjustment will also increase, for given values of β, the lower the ϵ, that is, the higher the share of nontraded goods in expenditures, λ, or the higher the elasticity of relative prices with respect to monetary imbalance, θ. The more rapid adjustment of the rate of exchange when ϵ falls (which also implies a faster acceleration of the rate of inflation although a smaller loss of reserves during the adjustment process) is illustrated by the comparison of paths II and I in figure 1.

Thus the results indicate that under a crawling-peg system a small country can choose its own long-run rate of inflation independently from the rest of the world. However, balance-of-payments deficits and surpluses are experienced in the process of adjustment, owing to purchasing power disparities which arise because the exchange rate takes some time to adjust fully.

We turn now to the analysis of the empirical results obtained by applying the model described above to the case of Brazil.

[15] That is, the difference between the new rate of ex ante excess supply of money $[\gamma D^* + \Delta(\gamma D^*) - m_d^*]$ and the world rate of inflation.

[16] For presentational simplicity the figure does not consider the effects of the temporary reduction in m_d^* as inflationary expectations accelerate. The pattern of adjustment is, however, very similar when these effects are taken into account.

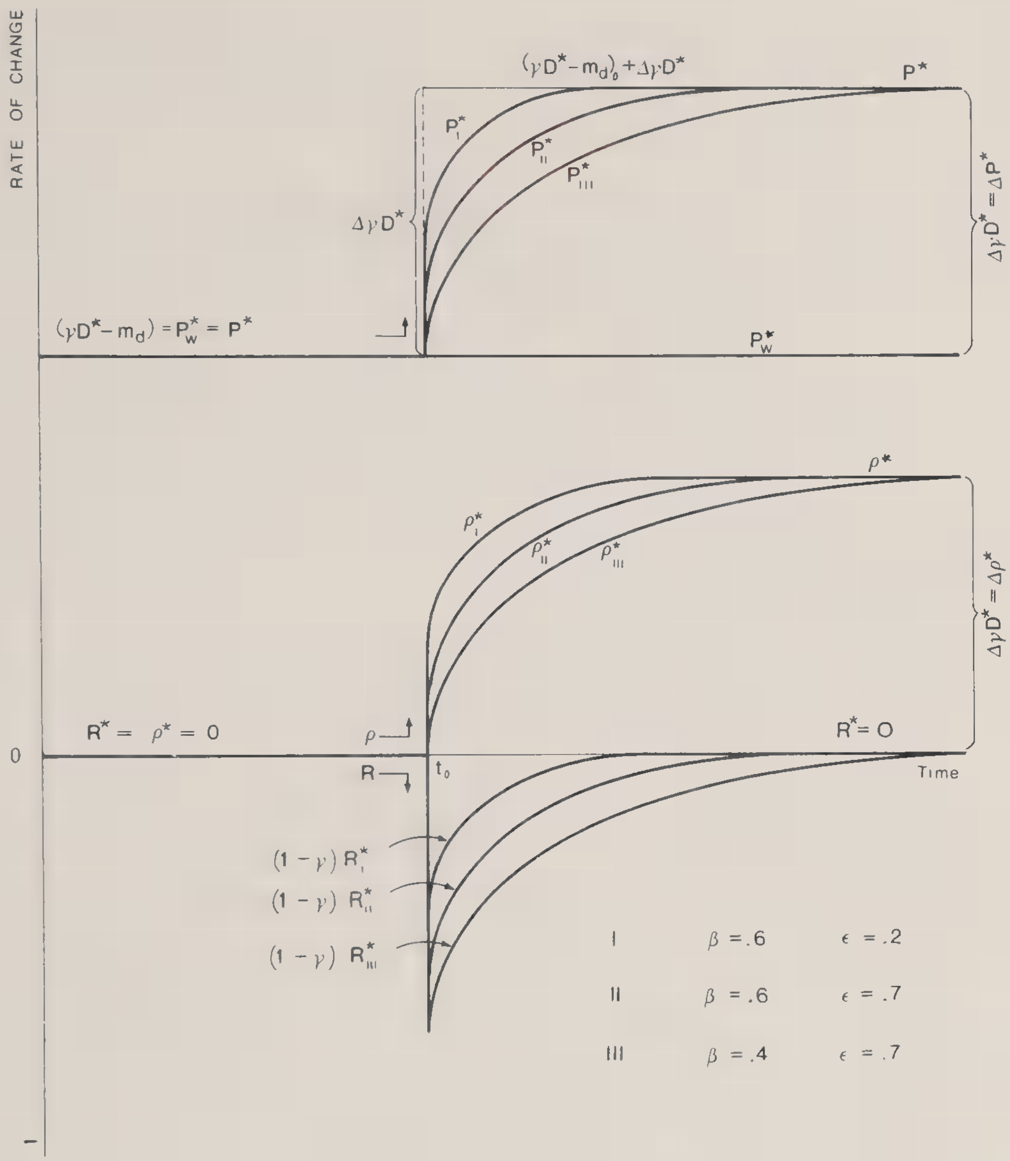


FIG. 1

II. Empirical Investigation

Estimation Procedures

In this section, the model developed above is restated in econometric form. For estimation purposes, we work with the following system:

$$\log m_t = \delta_1 \log y_t - \delta_2 \pi^e_t + \delta_3 \log m_{t-1} + \mu_{1t}, \tag{3'}$$

$$P^*_t = \epsilon(P^*_w + \rho^*)_t + (1 - \epsilon)(a^* + \gamma D^* - m^*_d)_t + \mu_{2t}, \tag{11'}$$

$$(1 - \gamma)R^*_t = \epsilon(P^*_w + \rho^* + m^*_d - \gamma D^* - a^*)_t + \mu_{3t}, \tag{12'}$$

$$\rho^*_t = \beta(P^* - P^*_w)_t + \beta(1 - \beta)(P^* - P^*_w)_{t-1} \tag{13'}$$

$$+ \beta(1 - \beta)^2(P^* - P^*_w)_{t-2} + \mu_{4t}.$$

Equation (3′) is the empirical counterpart of the model's demand for real money balances. The specification is semilogarithmic, in that the log of real money balances depends on the log of real income and expected inflation. As the model is applied to quarterly data, we postulate a process of gradual adjustment of real balances to their optimum value, and that is the rationale for the appearance of lagged money balances in the equation.[17] Equations (11′)–(13′) are stochastic variants of equations (11)–(13) above.[18]

The following stochastic specification is adopted: It is assumed that μ_i has a first-order autoregressive representation $\mu_{it} = \phi_i \mu_{it-1} + v_{it}$ (for $i = 1, 2, 3, 4$), where ϕ_i is the autoregressive parameter, and v_{it} is the error term. Furthermore, we assume that the error vectors, $v_{1t}, v_{2t}, v_{3t}, v_{4t}$ (for $t = 1, \ldots, T$) have zero mean and a constant variance-covariance matrix Γ; also, they are serially uncorrelated and drawn from a multivariate normal distribution.

Equations (3′) and (11′)–(13′) constitute a simultaneous system, which is linear in the variables but nonlinear in (some of) the parameters. Given these characteristics, consistent estimation requires the use of some simultaneous-equation method. We have used a full-information maximum-likelihood (FIML) estimator; FIML utilizes all the a priori restrictions on the system to estimate the coefficients of interest (in our case, ϵ, β, ϕ_i, δ_i) by maximizing the model's likelihood function.[19]

To make the estimation operational, a time series of m^*_{dt}, the rate of change of real money demand, is required. As money demand is simultaneously estimated with the other equations of the model, the complete system is estimated under the restriction that m^*_{dt} corresponds to the first difference of the fitted values of equation (3′). In addition, note that the estimation of money demand requires a proxy for π^e. This proxy is constructed by assuming a version of the rational expectations hypothesis. Specifically, it is postulated that agents form their inflation expectations on the basis of the least-squares prediction of inflation, conditional on the set of available information. We have

[17] Formally, the demand for money is given by $\log m_{d_t} = \delta_1' \log y_t - \delta_2' \pi^e_t$. Money demand equations of this form have previously been estimated for Brazil by Silveira (1973) and Khan (1977). We assume an adjustment mechanism of the form ($\log m_t - \log m_{t-1}) = h(\log m_{dt} - \log m_{t-1})$, $0 < h < 1$. See Griliches (1967) for a discussion of the conditions under which adjustment mechanisms of this class are likely to be optimal.

[18] As can be seen, we are now taking into account changes in the money multiplier. Note also that we consider only three terms in the exchange-rate equation (this issue is discussed further below).

[19] See Hendry (1971). The parameter estimates reported below were obtained by using Wymer's (1978) computer program, RESIMUL. The program uses a Newton-Raphson iterative procedure, beginning with arbitrarily given initial values of the parameters.

assumed that the relevant information set includes lagged values of the variables which, according to the model, determine the rate of inflation.[20]

A Crawling-Peg Experience: The Case of Brazil

Brazil's adoption of the crawling-peg regime can be considered as an extension of indexation to the external sector. Starting in 1964, Brazil engaged in one of the most comprehensive indexation programs anywhere. The financial sector was the first to be indexed. Exchange-rate indexation was introduced in August 1968.[21]

During the postwar period, Brazil had experienced a variety of exchange-rate regimes. Until 1953, exchange controls and licensing were widely used. In 1953 an exchange auction system with multiple exchange rates was introduced. A process of unifying the rates was initiated in 1961, and since then the country has had a pegged exchange rate. Since 1968, when the crawling peg was adopted, new selling and buying rates for the cruzeiro in relation to the U.S. dollar are announced by the central bank at frequent intervals (on average, the rate was changed over 10 times a year).

Despite the fact that there are no explicitly stated rules for the crawl, it is evident that the method in effect applies a purchasing power parity clause to the domestic price of foreign exchange, and the evolution of both domestic and foreign price levels is taken into account in adjusting the exchange rate. In this way, as emphasized by the Brazilian authorities, it is possible to neutralize the harmful effects of domestic inflation on the competitiveness of Brazilian products in international markets and, hence, on Brazil's balance of payments.[22]

Although inflation differentials are certainly an important factor in determining the crawl, it was not the only one. However, it is evident from the data (see Lemgruber 1977, tables 1 and 6; and Moura da Silva 1977, table 2) that since 1968 the variations in the exchange rate have followed purchasing power parity rates closely, suggesting that exchange-rate changes have been mainly determined by the differ-

[20] Specifically, π^e was constructed from the fitted values of a regression of the rate of inflation on lagged values of itself, foreign inflation, domestic-credit growth, exchange-rate devaluations, and real income growth. For each of these variables two lagged values were included in the equation.

[21] On the various aspects of indexation in Brazil, see Nadiri and Pastore (1977). A comprehensive study of the recent Brazilian economic experience is presented by Lemgruber (1977). On the functioning of the crawling peg and its effects on the economy, see Moura da Silva (1977). See also Lara-Resende (1978).

[22] This appears to be the main justification put forward by the authorities for the adoption of the crawling peg; on this issue see Lemgruber (1977) and Moura da Silva (1977).

ence between domestic and foreign inflation.[23] The specification of the exchange-rate reaction function above (eq. [13]) is intended precisely to capture this feedback process. Such a specification imposes restrictions on the data; econometric tests of these restrictions as well as the results from estimating the model are reported below.

Empirical Results

We now report the estimates of the parameters of equations (3′) and (11′)–(13′), as well as of the derived reduced-form equations. All estimates are based on quarterly data for Brazil. The sample period, III/1968 through IV/1977, contains 38 observations. Most of the variables are represented by readily available series.[24] The results are presented in table 1. All the parameter estimates have the proper sign and are reasonably well determined; in most cases asymptotic t ratios are well above 2.[25]

All the coefficients in the estimated money-demand equation (3′) have the expected signs and are significantly different from zero. The estimated coefficients are fairly similar to those reported for Brazil by Silveira (1973) and Khan (1977). Our parameter estimates imply a long-run income elasticity of real money demand of 1.307 and a long-run elasticity or real money demand with respect to expected inflation of −0.77.

The results for (11′)–(13′) indicate that both foreign inflation and excess supply in the money market significantly affect the determination of the domestic rate of inflation. These two factors, world inflation and monetary imbalance, also significantly affect the balance of payments, as shown by equation (12′). Specifically, an acceleration in the rate of domestic-credit creation over and above the rate of growth in money demand will increase domestic inflation and reduce international reserves. The overall effect of such an increase, however, cannot be established solely from equations (11′) and (12′). This

[23] The exchange rule has been followed with a considerable degree of consistency. Almonacid and Pastore (1977) present some evidence to that effect and criticize the lack of flexibility—the rule has not been altered to deal with the oil crisis and with the mounting external debt of Brazil. Similar evidence is also presented by Connolly and da Silveira (1979, pt. 3), who found that purchasing power parity holds very well, particularly after 1962.

[24] The following series were used for P, R, and M: consumer price index, international reserves of the central bank, and the money supply; D was generated by subtracting R from the monetary base; ρ is the market exchange rate, selling rate (average over the quarter) (all the above were taken from *International Financial Statistics*, various issues); y is represented by the quarterly real income series constructed by Wachter (1976). (The series were extended through 1977 using Wachter's method.)

[25] Asymptotic t ratios are defined as coefficient values divided by asymptotic standard errors.

TABLE 1

FULL-INFORMATION MAXIMUM-LIKELIHOOD ESTIMATES OF THE MODEL

(3′) $$\log m_t = \underset{(0.049)}{0.162} \log y_t - \underset{(0.229)}{1.700}\, \pi_t^e + \underset{(0.045)}{0.876} \log m_{t-1},$$

$$\text{MSE} = 0.0018,\ \phi_1 = \underset{(0.042)}{-0.613},\ C_1 = \underset{(0.297)}{0.873};$$

(11′) $$P_t^* = \underset{(0.018)}{0.958}(P_w^* + \rho^*)_t + \underset{(0.018)}{0.042}(a^* + \gamma D^* - m_d^*)_t,$$

$$\text{MSE} = 0.0003,\ \phi_2 = \underset{(0.110)}{-0.082},\ C_2 = \underset{(0.0012)}{0.002};$$

(12′) $$(1-\gamma)R_t^* = \underset{(0.018)}{0.958}(P_w^* + \rho^* + m_d^* - \gamma D^* - a^*)_t,$$

$$\text{MSE} = 0.0021,\ \phi_3 = \underset{(0.045)}{0.251},\ C_3 = \underset{(0.039)}{-0.038};$$

(13′) $$\rho_t^* = \underset{(0.106)}{0.526}(P^* - P_w^*)_t + \underset{(0.005)}{0.249}(P^* - P_w^*)_{t-1} + \underset{(0.029)}{0.118}(P^* - P_w^*)_{t-2},$$

$$\text{MSE} = 0.0005,\ \phi_4 = \underset{(0.106)}{-0.050},\ C_4 = \underset{(0.003)}{-0.008};$$

$$\Gamma = \begin{bmatrix} 0.0018 & & & \\ -0.0004 & 0.0006 & & \\ 0.0009 & 0.0004 & 0.0029 & \\ 0.0002 & -0.0005 & -0.0005 & 0.0005 \end{bmatrix}.$$

NOTE.—See text for notation. Numerals in parentheses are asymptotic standard errors of regression coefficients; Γ is the estimated variance-covariance matrix of residuals; C_i (i = 1, 2, 3, 4) is the estimated constant term; ϕ_i is the autoregression coefficient; and MSE is the mean square error.

is so because the divergence of domestic from world inflation will, as indicated by equation (13′), depreciate the exchange, which in turn will further increase the rate of inflation and reduce the rate at which foreign reserves are depleted. In order to analyze the full effect of changes in the rate of growth of domestic credit (as well as in the other exogenous variables), reduced-form coefficients should be considered. These coefficients are reported below.

The parameter estimates in equation (13′)—the exchange-rate reaction function—indicate that more than 50 percent of the foreign-domestic inflation differential is transmitted to the exchange rate within the current quarter, and an additional 25 percent in the next. After only three quarters the full differential is completely reflected by the exchange-rate adjustment. In this context, it is important to note that equation (13) embodies a restriction regarding the effects of current and lagged values of the domestic-foreign inflation differential on the exchange rate. Specifically, a pattern of geometrically decaying weights was postulated. The results reported in table 1 are derived under the imposition of this restriction. To test

TABLE 2

DERIVED REDUCED-FORM ESTIMATES

(14′) $$\rho_t^* = \underset{(0.018)}{(0.045} + \underset{(0.017)}{0.43L} + \underset{(0.015)}{0.41L^2} + \underset{(0.014)}{0.039L^3} + \underset{(0.012)}{0.037L^4)}(\gamma D^* + a^* - m_d^* - P_w^*)_t,$$

(15′) $$P_t^* = \underset{(0.030)}{(0.915} - \underset{(0.016)}{0.041L} - \underset{(0.014)}{0.039L^2} - \underset{(0.013)}{0.037L^3} - \underset{(0.012)}{0.036L^4)}(P_w^*)_t + \underset{(0.030)}{(0.085} + \underset{(0.016)}{0.041L} + \underset{(0.014)}{0.039L^2} + \underset{(0.013)}{0.037L^3} + \underset{(0.012)}{0.036L^4)}(\gamma D^* + a^* - m_d^*)_t,$$

(16′) $$(1-\gamma)R_t^* = \underset{(0.030)}{(-0.915} + \underset{(0.016)}{0.041L} + \underset{(0.014)}{0.039L^2} + \underset{(0.013)}{0.037L^3} + \underset{(0.012)}{0.036L^4)}(\gamma D^* + a^* - m_d^* - P_w^*)_t.$$

NOTE.—For notation and explanation, see text.

its empirical validity, we have estimated the system unrestricted, and a χ^2 likelihood ratio test was constructed, giving a value of 2.114 with 2 df. This is less than the 95 percent χ^2 critical value; that is, the sample information fails to reject the restrictions embodied in (13′) at the usual 5 percent significance level.

With the information contained in table 1 it is possible to calculate the derived reduced-form estimates and their asymptotic standard errors. The model's reduced forms are given by equations (14)–(16), whose estimation forms include changes in the money multiplier, a^*. These estimates, as implied by the structural estimates of table 1, are presented in table 2 ([14′]–[16′]), from which the overall effects of current and lagged rates of excess money supply and of foreign inflation on the exchange rate, the domestic rate of inflation, and the balance of payments can be assessed.[26] The coefficients indicate that about 25 percent of the excess flow supply of money (the difference between the rates of growth of domestic credit and money demand) is transmitted to the inflation rate within five quarters. To the extent that the rate at which the excess supply of money is created exceeds the foreign rate of inflation, the exchange rate will depreciate (by about 15 percent of the difference, during the first five quarters) and international reserves will fall. Monetary equilibrium is therefore maintained by a combination of reductions in the real value of the nominal stock (due to the acceleration in the inflation rate) and reductions in the monetary base (through the loss of reserves).

The effects of foreign inflation on the three endogenous variables can also be analyzed from the estimates of table 2. A higher rate of world inflation will reduce the rate of exchange depreciation and will

[26] Although only current and four lagged coefficients are reported, the coefficients for additional lags can be calculated from the information provided in table 1.

have a positive effect on the rate of accumulation of foreign reserves. It will, however, strongly and significantly raise the domestic rate of inflation during the current quarter. Note that lagged rates of foreign inflation appear with a minus sign, which is indeed consistent with our version of the crawling-peg regime. An acceleration of world inflation will, ceteris paribus, induce subsequent appreciations of the exchange rate, thus moderating the initial increase in both the domestic price of traded goods and the domestic rate of inflation.

III. Concluding Remarks

The purpose of this study has been to construct and test a model that accounts for the joint determination of the exchange rate, the rate of inflation, and the balance of payments in a crawling-peg economy. The model presented extends the previous literature on the monetary approach to the case of a crawling-peg regime.

According to the model, a small open economy that indexes its exchange rate through the adoption of a purchasing power parity clause may choose its rate of inflation independently from the rest of the world. To the extent that purchasing power disparities arise because full adjustment of the exchange rate is not instantaneous, equilibrating flows of international reserves will take place. Thus the model is capable of generating a pattern of short-run deviations from purchasing power parity which occur simultaneously with movements in international reserves and the exchange rate.

Our theoretical framework can be used to analyze the effects of domestic monetary policy and of external inflation on a crawling-peg economy. For example, we have shown that an increase in the rate of domestic-credit creation will generally tend to raise domestic inflation, reduce international reserves, and depreciate the exchange. The exact path of adjustment will depend on a number of parameters explicitly incorporated into the model. The faster the adjustment of the exchange rate to purchasing power disparities, for example, the greater the impact effects of domestic monetary variables on the rate of inflation and the smaller their impact effect on the balance of payments. Similarly, domestic monetary variables will have a greater impact effect on domestic inflation, and a smaller impact effect on the balance of payments, the higher the share of nontraded goods in expenditures and the higher the elasticity of the traded/nontraded relative price with respect to these monetary variables.

The predictions of the model regarding signs and magnitudes of the different parameters appear to be sustained by the Brazilian data. Overall, the empirical evidence presented seems to indicate that our theoretical framework is consistent with the sample information.

Appendix

In this Appendix we derive the structural and reduced-form equations of the model when persistence effects are allowed in the price equation for nontraded goods. The structural equations using this specification are then estimated and the results are reported below.

Replacing equation (8) by equation (9), we can express the rate of change of nontraded goods prices as

$$(P^*_{\mathrm{NT}})_t = (P^*_{\mathrm{T}})_t + \frac{\theta}{1 - \alpha L}\Omega_t. \tag{A1}$$

Adopting now the definition of Ω_t given in Section I, we obtain equation (A2), the counterpart of equation (10),

$$(P^*_{\mathrm{NT}})_t = (P^*_{\mathrm{T}})_t + \frac{\theta}{1 - \alpha L}(\gamma D^* + a^* - M^*_d)_t. \tag{A2}$$

Using equations (7) and (A2), the rate of domestic inflation, equation (6), can be written as

$$P^*_t = \epsilon'(P^*_w + \rho)_t + (1 - \epsilon')(\gamma D^* + a^* - m^*_d)_t, \tag{A3}$$

where ϵ' is now a polynomial in the lag operator,

$$\epsilon' = \frac{1 - \alpha L}{1 - \alpha L + \theta(1 - \lambda)}. \tag{A4}$$

Postulating the same crawling rule as used in the text, that is, equation (13), we can obtain the reduced forms of the model,[27]

$$\rho^*_t = \frac{\beta(1 - \epsilon')}{1 - \beta\epsilon' - (1 - \beta)L}(\gamma D^* + a^* - m^*_d - P^*_w)_t, \tag{A5}$$

$$P^*_t = \frac{\epsilon'(1 - \beta)(1 - L)}{1 - \beta\epsilon' - (1 - \beta)L}(P^*_w)_t + \frac{(1 - \epsilon')[1 - (1 - \beta)L]}{1 - \beta\epsilon' - (1 - \beta)L}(\gamma D^* + a^* - m^*_d)_t, \tag{A6}$$

$$(1 - \gamma)R^*_t = \frac{\epsilon'(1 - \beta)(1 - L)}{1 - \beta\epsilon' - (1 - \beta)L}(P^*_w + m^*_d - \gamma D^*)_t. \tag{A7}$$

The econometric forms of the model, in conjunction with the demand for money, equation (3′), are estimated by the procedure described in Section II. The parameter estimates do not differ much from the model in which this type of persistence is ignored.

For comparison and completeness we now report the estimates obtained with the extended model:

[27] The reduced forms could be expressed in terms of the underlying parameters, λ, α, and θ by substituting the RHS of (A4) for ϵ'. This results in rather cumbersome expressions, as, e.g., in eq. (A5):

$$\rho^*_t = \frac{\beta\theta(1 - \lambda)}{(1 - \alpha L)(1 - \beta)(1 - L) + \theta(1 - \lambda)(1 - \beta)L}(\gamma D^* + a^* - m^*_d - P^*_w)_t.$$

$$\log m_t = \underset{(0.063)}{0.224} \log y_t - \underset{(0.300)}{2.224}\pi_t^e + \underset{(0.058)}{0.828} \log m_{t-1}, \tag{A3$'$}$$

$$\rho_t^* = \underset{(0.099)}{0.618}(P^* - P_w^*)_t + \underset{(0.023)}{0.236}(P^* - P_w^*)_{t-1} + \underset{(0.032)}{0.090}(P^* - P_w^*)_{t-2}, \tag{A8}$$

$$P_t^* = \underset{(0.018)}{0.951}(P_w^* + \rho^*)_t + \underset{(0.018)}{0.049}(a^* + \gamma D^* - m_d^*)_t, \tag{A9}$$

$$(1 - \gamma)R_t^* = \underset{(0.018)}{0.951}(P_w^* + \rho^* + m_d^* - \gamma D^* - a^*)_t. \tag{A10}$$

References

Almonacid, Rubén D., and Pastore, Alfonso C. "El tipo de cambio, la crisis del petroleo y la deuda externa de Brasil." *Cuadernos econ.* (Santiago, Chile) 14, no. 43 (December 1977): 109–27.

Artus, Jacques R. "Exchange Rate Stability and Managed Floating: The Experience of the Federal Republic of Germany." *International Monetary Fund Staff Papers* 23 (July 1976): 312–33.

Blejer, Mario I. "The Short-Run Dynamics of Prices and the Balance of Payments." *A.E.R.* 67 (June 1977): 419–28.

———. "On Causality and the Monetary Approach to the Balance of Payments: The European Experience." *European Econ. Rev.* 12 (July 1979): 289–96.

Boyer, Russell S. "Optimal Foreign Exchange Market Intervention." *J.P.E.* 86, no. 6 (December 1978): 1045–55.

Connolly, Michael, and da Silveira, José Dantas. "Exchange Market Pressure in Postwar Brazil: An Application of the Girton-Roper Monetary Model." *A.E.R.* 69 (June 1979): 448–54.

Cooper, Richard N. "Sliding Parities: A Proposal for Presumptive Rules." In *Approaches to Greater Flexibility of Exchange Rates: The Bürgenstock Papers*, edited by George N. Halm. Princeton, N.J.: Princeton Univ. Press, 1970.

Dean, James W. "Problems in the Specification and Interpretation of Central Bank Reaction Functions." *Econ. and Soc. Rev.* 5 (July 1974): 431–43.

Dornbusch, Rudiger. "Devaluation, Money, and Nontraded Goods." *A.E.R.* 63 (December 1973): 871–80.

———. "The Theory of Flexible Exchange Rate Regimes and Macroeconomic Policy." *Scandinavian J. Econ.* 78 (May 1976): 255–75.

Frenkel, Jacob A. "A Monetary Approach to the Exchange Rate: Doctrinal Aspects and Empirical Evidence." *Scandinavian J. Econ.* 78 (May 1976): 200–24.

———. "The Demand for International Reserves under Pegged and Flexible Exchange Rate Regimes and Aspects of the Economics of Managed Float." In *The Functioning of Floating Exchange Rates: Theory, Evidence, and Policy Implications*, edited by David Bigman and Teizo Taya. Cambridge, Mass.: Ballinger, 1980.

Frenkel, Jacob A., and Clements, Kenneth W. "Exchange Rates in the 1920's: A Monetary Approach." In *Development in an Inflationary World*, edited by M. J. Flanders and Assaf Razin. New York: Academic Press, in press.

Frenkel, Jacob A., and Johnson, Harry G., eds. *The Monetary Approach to the*

Balance of Payments. London: Allen & Unwin; Toronto: Univ. Toronto Press, 1976.

———, eds. *The Economics of Exchange Rates: Selected Studies*. Reading, Mass.: Addison-Wesley, 1978.

Friedlaender, Ann F. "Macro Policy Goals in the Postwar Period: A Study in Revealed Preference." *Q.J.E.* 87 (February 1973): 25–43.

Girton, Lance, and Roper, Don. "A Monetary Model of Exchange Market Pressure Applied to the Postwar Canadian Experience." *A.E.R.* 67 (September 1977): 537–48.

Griliches, Zvi. "Distributed Lags: A Survey." *Econometrica* 35 (January 1967): 16–49.

Hahn, Frank H. "The Monetary Approach to the Balance of Payments." *J. Internat. Econ.* 7 (August 1977): 231–49.

Hendry, David F. "Maximum Likelihood Estimation of Systems of Simultaneous Regression Equations with Errors Generated by a Vector Autoregressive Process." *Internat. Econ. Rev.* 12 (June 1971): 257–72.

International Monetary Fund (IMF). *The Monetary Approach to the Balance of Payments*. Washington: IMF, 1977.

Johnson, Harry G. "The Monetary Approach to the Balance-of-Payments Theory." *Further Essays in Monetary Theory*. London: Allen & Unwin, 1972.

———. "The Monetary Approach to the Balance of Payments: A Nontechnical Guide." *J. Internat. Econ.* 7 (August 1977): 251–68.

Kenen, Peter B. "Floats, Glides and Indicators: A Comparison of Methods for Changing Exchange Rates." *J. Internat. Econ.* 5 (May 1975): 107–51.

Khan, Mohsin S. "Variable Expectations and the Demand for Money in High-Inflation Countries." *Manchester School Econ. and Soc. Studies* 45 (September 1977): 270–93.

Kouri, Pentti J. K. "Exchange Rate Expectations, and the Short-Run and the Long-Run Effects of Fiscal and Monetary Policies under Flexible Exchange Rates." Paper presented at the conference on Monetary Mechanisms in Open Economics, Helsinki, August 1975.

Lara-Resende, Andre. "Interest Parity and the Brazilian Crawling Peg: The Official and Parallel Markets." Mimeographed. Cambridge, Mass.: Massachusetts Inst. Technol., 1978.

Leiderman, Leonardo. "Relationships between Macroeconomic Time-Series in a Fixed-Exchange-Rate Economy." *European Econ. Rev.* 14 (July 1980): 61–77.

Lemgruber, Antonio C. "Inflation in Brazil." In *Worldwide Inflation: Theory and Recent Evidence*, edited by Lawrence B. Krause and Walter S. Salant. Washington: Brookings Inst., 1977.

Levin, Jay H. "Monetary Policy and the Crawling Peg." *Econ. J.* 85 (March 1975): 20–32.

———. "Reserve Stocks as External Targets and the Stability of Alternative Exchange Rate Systems." *Rev. Econ. Studies* 44 (February 1977): 59–70.

Magee, Stephen P. "The Empirical Evidence on the Monetary Approach to the Balance of Payments and Exchange Rates." *A.E.R. Papers and Proc.* 66 (May 1976): 163–70.

Mathieson, Donald J. "Is There an Optimal Crawl?" *J. Internat. Econ.* 6 (May 1976): 183–202.

Moura da Silva, Adroaldo. "The Basis of the Minidevaluation Policy." *Explorations Econ. Res.* 4 (Winter 1977): 101–16.

Mundell, Robert A. *International Economics*. New York: Macmillan, 1968.

———. *Monetary Theory: Inflation, Interest, and Growth in the World Economy.* Pacific Palisades, Calif.: Goodyear, 1971.

Mussa, Michael. "The Exchange Rate, the Balance of Payments, and Monetary and Fiscal Policy under a Regime of Controlled Floating." *Scandinavian J. Econ.* 78 (May 1976): 229–48.

Nadiri, M. Ishaq, and Pastore, Alfonso C., eds. *Indexation: The Brazilian Experience.* Special issue of *Explorations Econ. Res.*, vol. 4 (Winter 1977).

Parkin, Michael. "Inflation, the Balance of Payments, Domestic Credit Expansion, and Exchange Rate Adjustments." In *National Monetary Policies and the International Financial System*, edited by Robert Z. Aliber. Chicago: Univ. Chicago Press, 1974.

Silveira, Antonio M. "The Demand for Money: The Evidence from the Brazilian Economy." *J. Money, Credit and Banking* 5 (February 1973): 113–40.

Wachter, Susan M. *Latin American Inflation: The Structuralist-Monetarist Debate.* Lexington, Mass.: Heath, 1976.

Whitman, Marina V. N. "Global Monetarism and the Monetary Approach to the Balance of Payments." *Brookings Papers Econ. Activity*, no. 3 (1975), pp. 491–536.

Williamson, John H. *The Crawling Peg.* Princeton Essays in International Finance no. 50. Princeton, N.J.: Princeton Univ. Press, 1966.

Wymer, C. *Computer Programs: Resimul Manual.* Washington: IMF, 1978.

Does Federalism Matter? Political Choice in a Federal Republic

Susan Rose-Ackerman

Yale University

This paper builds upon some well-known facts about state government to generate new conclusions about social choice on the national level of a federal republic. Citizens vote against national laws that restrict their state's ability to export costs but support laws that reduce the costs imposed on them. Individuals may seek to extend the laws passed in some states to the entire nation or may oppose preemptive laws because they benefit from variety. Since these motivations are absent in a unitary system, national support for a law will depend upon whether a unitary or a federal structure prevails.

I. Introduction

This paper builds upon some well-known facts about state government to generate new conclusions about social choice on the national level of a federal republic. The central feature of state government behavior I shall exploit is the incentive each state has to improve its own position by imposing costs on the residents of other states. This search for local advantage takes many forms. On the one hand, states may impose taxes and regulations which are borne by out-of-state residents; on the other hand, they may try to attract investment from other states by providing tax breaks and special public services.[1]

This paper was presented at the Workshop on Analytical Urban Economics at Queen's University, Kingston, June 1978. I wish to thank Bruce Ackerman, the workshop participants, and the referee of this *Journal* for helpful comments.

[1] These incentives are stressed by Walker (1969, p. 890) and Posner (1977, chap. 26). McLure (1967) estimates that about one-quarter of all state taxes are exported. Externalities such as air and water pollution are emphasized by Breton (1965), Olson (1969), Tullock (1969), Rothenberg (1970), and Oates (1972). These authors discuss how a federal system can trade off interjurisdictional variety against the internalization of externalities through boundary definition, the assignment of functions to levels of government, and intergovernmental grants.

[*Journal of Political Economy*, 1981, vol. 89, no. 1]

Whatever their particular character, these interstate spill-ins and spill-outs can alter the substance of national legislation. Citizens will vote against laws that restrict their state's ability to export costs but will support laws that reduce the costs imposed on them by their own and other states' choices. Since these strategic motivations are absent in a unitary system, national support for particular laws will depend upon whether a unitary or a federal structure prevails.

Despite its potential importance, the impact of federalism on the strategic position of voters has been largely ignored by social scientists. In his excellent review of the literature, for example, William Riker (1975) argues that federalism has no important influence on substantive national outcomes, suggesting that it does no more than delay the passage of national legislation opposed by a substantial minority of state legislatures.[2] Although Riker's conclusion is partly an empirical proposition, it is also based on an implicit model of how government systems behave. Riker argues (1975, pp. 155–56) that the underlying distribution of tastes in the population determines political outcomes. Since federalism per se has no impact on tastes, it will not, therefore, affect national political choices in the long run. The present paper challenges these theoretical underpinnings.

Since I am concerned not with an exhaustive taxonomy but with the relationship between political power and political structure, I concentrate on simple models. The system I discuss has only two "layers"—national and state.[3] States have fixed boundaries and cover the nation's entire geographical area so that every citizen lives in only one state. I will contrast the legislative choices of a unitary government with those of a "hierarchical" federal system where higher-level governments can always preempt the legislative choices of lower-level governments. If the superior government has taken no affirmative action, however, the statutes of inferior governments are binding.[4] I

[2] Riker (1975) proposes an imaginary experiment in which matched pairs of unitary and federal systems are examined to discover if there are important public policy differences. He hypothesizes that public policy within each pair will be "remarkably similar regardless of federalism" (p. 144). Furthermore, he argues that in the United States federalism delayed national regulation of business (p. 154) and helped perpetuate racist acts (pp. 154–56) but had no long-run impact on the character of national legislation.

[3] For an attempt to justify a three-layered system, see Ylvisaker (1959). With the exception of Wechsler (1963), Posner (1977), and Winter (1977), legal commentators have all but ignored the issues discussed in this paper and have instead concentrated on the relationship between state and federal courts. For a recent article in this tradition, see Cover and Aleinikoff (1977).

[4] This model should be contrasted with others which have a strict division of functions between high- and low-level governments. For example, see Wheare (1953, pp. 32–33). Recently, concern for a strict division of authority has given way to scholarship which recognizes the importance of interactions between levels of government. See, for example, Grodzins (1960, 1966) and Elazar (1962).

assume that both state and national governments are direct democracies where the only political actors are individuals and where all decisions are made by majority vote. Political issues are separable and well defined, and they appear to voters as simple dichotomous choices between maintenance of the status quo and a change.[5] The analysis concentrates on legislative choices. I do not discuss administrative or judicial issues or the possibility that the national government will seek to administer programs through state government agencies.

Section II shows that federalism "matters" in a political system where capital can move but people cannot. Section III considers the impact of permitting interstate migration of voters, and Section IV shows that federalism can matter even in the long run when states respond to the legislative choices of other states.

II. Federalism Does Matter

The major difference between a unitary system and a hierarchical federal structure is the characterization of the status quo. Let y = the status quo in a unitary system, and let x = a proposed, exogenously defined, legislative change. Suppose, for example, that in the status quo casino gambling is illegal. The federal government, however, proposes to make such gambling legal and to levy a tax on the profits.

In a hierarchical federal system, some states have passed legislation that is similar to x. Others have a status quo that is similar to y. Let us call the former x-type states and the latter y-type states. Continuing the example, x-type states permit casino gambling and levy a tax on earnings, while casino gambling is illegal in y-type states. Assume that if all states are x-type states, then the state laws taken together would have the same impact as x. Similarly, assume that if no states have passed the x-type law, the status quo is the same as in a unitary system. Thus, in my example the state gambling laws are duplicates of x and y, respectively. This assumption is not essential to the analysis, but it simplifies the exposition considerably.

Because of interstate spillovers of costs and benefits, individuals' preferences over x and y will not necessarily be the same as their preferences for alternative state legal regimes. In the gambling example, some people may favor a statewide gambling law and vote

[5] These assumptions permit me to look at the vote on a single issue and to avoid voting cycles in which majority rule does not produce a determinant outcome (Sen 1970, p. 38). The actual relationship between state statutes and federal law in the same area is, of course, considerably more complex. In reality, the relationship is seldom obvious and federal courts are frequently called upon to sort out the overlapping authority of state and federal statutes (see Tribe 1978, pp. 378–86).

against x at the national level. A state's residents gain when out-of-staters come to gamble and consume tourist services. The state would lose this advantage if gambling were legal in every state. In a unitary system, however, these same individuals might favor x over y since they would have no special advantage to protect.

A voter's preference for x over a system with a variety of state laws will often depend upon the number and location of x-type states. The economic advantages of permitting gambling in one's state of residence are larger the smaller the number of other states which also permit gambling. The benefits are also higher the more geographically distant are the other jurisdictions with legalized gambling. Let $g_i = z_i$ if state i has passed an x-type law, $g_i = w_i$ if state i is a y-type state, and let H = total number of states. Then, the status quo in a federal system is a vector $G = (g_1, \ldots, g_i, \ldots, g_H)$ where $g_i = z_i$ or w_i. Thus we can summarize the situation of those living in x-type states as $z(G)$ and the situation of those in y-type states as $w(G)$.

In this portion of the analysis, I assume that migration across state lines is impossible, so that the only way individuals can affect the legal regime under which they live is to vote for new state or federal laws. Capital, however, is free to move between jurisdictions. Given these assumptions, we can now compare the way voters with different preferences will cast their ballots in federal and unitary republics. Turning first to a unitary system, individual preferences can be straightforwardly translated in a social choice. Law x passes if and only if:

$$N(xRy) > N(yRx), \tag{1}$$

where R is the binary relation of "weak preference" ("at least as good as"). Then $N(aRb)$ is the number of people for whom aR_jb where R_j is the weak preference relation for individual j and R_j is reflexive, transitive, and complete.[6]

In a hierarchical federal system, under the assumption that only the current level of G is relevant, x passes at the national level if and only if:

$$N_1[xRz(G)] + N_2[xRw(G)] \geqslant N_1[z(G)Rx] + N_2[w(G)Rx], \tag{2}$$

where $N_1(aRb)$ is the number of people living in x-type states for whom aR_jb and $N_2(aRb)$ is the number of people living in y-type states for whom aR_jb.

Expressions (1) and (2) need not yield the same results. Individuals who would prefer x to y need not also prefer x to $z(G)$ or $w(G)$. If a

[6] In Sen's terminology (1970, p. 9), R_j is an ordering. The notation in (1) follows Sen (1970, p. 71).

TABLE 1

	$xRz(G)$	$z(G)Rx$	$xRw(G)$	$w(G)Rx$
xRy	A_1 (favor x)	A_3 (positive spill-ins to x types)	B_1 (favor x)	B_3 (positive spill-ins to y types)
yRx	A_2 (negative spill-ins to x types)	A_4 (oppose x)	B_2 (negative spill-ins to y types)	B_4 (oppose x)

state is able to export the costs of a program, then its citizens are likely to oppose preemptive federal laws even though they favor x-type legislation in their own state and would favor x over y in a unitary system. Similarly, if citizens can benefit from the legislative initiatives of other states then they may vote against x although they have xR_jy with a unitary government. The gambling example falls in the first category. Public health or pollution control laws in neighboring states may produce the second voting pattern. In fact, individual preferences may produce any of the possible rankings of x, y, and $z(G)$ and x, y, and $w(G)$; and one can thus divide the population into groups depending upon their preferences and the states in which they live. For a given G, let $A(G)$ = set of people in x-type states and let $B(G)$ = everyone else.

People living in x-type states have $xRz(G)$ or $z(G)Rx$ or both. Those in set $B(G)$ have $xRw(G)$ or $w(G)Rx$ or both. In a unitary system voters have xRy or yRx or both. Table 1 is a matrix of the possible taste combinations. The entries in the matrix are the sets of people in each taste class. Thus $A(G) = A_1(G) \cup A_2(G) \cup A_3(G) \cup A_4(G)$ and $B(G) = B_1(G) \cup B_2(G) \cup B_3(G) \cup B_4(G)$.[7] In expression (1), $N(xRy) = N(A_1 \cup A_3 \cup B_1 \cup B_3)$. In expression (2), $N_1[xRz(G)] = N_1(A_1 \cup A_2)$ and $N_2[xRw(G)] = N_2(B_1 \cup B_2)$. Thus given G, the vote for x in a unitary system, $N(xRy)$, is greater than, equal to, or less than the vote in a hierarchical system, $N_1[xRz(G)] + N_2[xRw(G)]$, as:

$$N(A_3 \cup B_3) \gtreqless N(A_2 \cup B_2). \tag{3}$$

Expression (3) shows that federalism "matters." The size of the vote in favor of the national law may be higher or lower in a hierarchical federal system than in a unitary system. A law which passes in one system may be defeated in another. The practical importance of this result, however, depends upon whether sets A_3, B_3 and A_2, B_2 represent common preference configurations. Sets A_3 and B_3 include indi-

[7] The symbol $\cup$ stands for the union of sets, i.e., the set of people belonging to either set.

viduals who prefer x to y as the national law in a unitary system but who each favor conditions in their own states over x.[8] Sets A_2 and B_2 are the reverse of A_3 and B_3. They include individuals who would oppose x in a unitary system but favor preemptive legislation because it keeps them from losing at the expense of other citizens.

If sets A_2 and B_2 are large relative to sets A_3 and B_3, then a paradoxical result is possible. Although federalism is often justified as a way to preserve diversity and prevent the centralization of power, (3) implies that the existence of low-level governments may cause some voters to favor central control of an activity that they would otherwise have wanted unregulated. Without constitutional limits, a majority of citizens may be so unhappy with the mixture of independent state choices under federalism that they favor a more powerful central government than they would in a unitary system.

III. Migration and the Vote for National Laws

When voters can move between jurisdictions in response to public policy changes, the vote for national preemptive laws will change. If migration is costless and if legal regimes are the only determinant of location, then, given G, national electoral support would fall in a hierarchical federal system. Voters are less likely to want national uniformity when they can move to congenial states.

Reality, of course, is somewhere between the extremes of a perfectly fixed population and a perfectly mobile one. Furthermore, different types of people have different moving costs, and the relation between these costs will help determine the political power of various groups at the state level. Thus, mobile groups may impose costs on immobile groups as a price for not leaving the state. Alternatively, undesirable groups may be induced to migrate by taxing or regulating them or by subsidizing their moving expenses. Variations in the cost of migration change the strategic environment. Those who can move cheaply have an advantage over those who cannot. Therefore, national decisions which require uniformity will be strongly supported by those who lose from the migration of others and opposed by those whose mobility permits them to gain at the expense of less mobile voters.

The possibility of migration has implications for national programs which seek to redistribute income from high-income to low-income families. Economic analyses of multiple government systems (e.g., Olson 1969; Oates 1972) conclude that interjurisdictional mobility

[8] It also, of course, includes those who are indifferent between any of these alternatives. Only people who are indifferent to all three alternatives are in both A_2 and A_3 or in both B_2 and B_3.

makes serious redistribution impossible at the state level. Progressive tax and spending policies must be carried out by preemptive laws at the national level. Interstate migration, however, also lowers the prospects for passage of national redistributive laws. If voters are mobile, support for a national preemptive law which redistributes income from the rich to the poor may be less than the support such legislation would obtain in a unitary system or in a federal system with no migration. Many who would support a particular progressive tax and spending program in a unitary system may oppose it in hierarchical federalism if the law preempts state "beggar-my-neighbor" laws which benefit these citizens. Selfishness will dominate altruism if the opportunity cost of altruism is too high.

IV. Does Federalism Matter in the Long Run?

The result in (3) applies when the number of x-type and y-type states is held constant (i.e., for a given G). One student of federalism (Riker 1975), however, argues that in the long run a federal system and a unitary one will converge to the same pattern of national laws. The view that federalism does not matter in the long run is based either on a hypothesis about how tastes change over time[9] or on a model that makes particular assumptions about the way votes change in response to new information and new circumstances (i.e., changes in G). The cases where federalism does not matter are contingent upon these assumptions. This section shows that it is possible to impose other plausible conditions on tastes, opportunities, and government structure and produce an equilibrium where x- and y-type states coexist in spite of majority support for x over y.

My analysis of state legislative choices challenges the idea that states are "lighthouses" that show the way to the federal government by enacting innovative laws. In my model, a law may spread to many states, not because it has been "tested" and found useful,[10] but rather

[9] Riker's argument turns on the overriding importance of individual tastes in determining political outcomes. He uses the example of white racism. "As long as whites strongly prefer racist institutions, one can expect institutions to be racist regardless of whether the country is federal or unitary. But when the preference for racist institutions weakens, then federalism helps racism by rendering difficult the enforcement of an anti-racist policy on the minority of white racists. So we can say that the beneficiaries of federalism get only marginal benefits on policy, but marginal or not, they are undoubtedly real" (1975, p. 156).

[10] See Walker (1969, 1971, 1973). By his choice of words, Walker appears to believe that early adopters of laws are progressive. He calls them "pioneers" (1969, p. 881). A glance at Walker's list of laws should warn anyone against this inference. Many involve the licensing of occupations such as barbers or real estate brokers. Others simply mandate the establishment of state agencies, many of which are required as a condition for receiving federal grants. For a different perspective on the same issue which emphasizes politicians' incentives to take risks, see Rose-Ackerman (1980).

because voters in y-type states want to avoid damage at the expense of others. Similarly, a national law may eventually pass in a federal system, not because a new initiative has been tested in the states, but rather because voters want to override the costs of spillovers and inconsistent laws.

Since a fully general analysis would be difficult to interpret, I concentrate on three special cases that capture the essential features of many actual situations. In these examples, I assume that tastes do not change over time, that voters have similar initial information, and that benefits that spill over to people in one group of states are costs to people in the other. In case 1, voters in x-type states receive positive spill-ins from out-of-state residents, and voters in y-type states bear costs. In case 2, voters in x-type states receive negative spill-ins from y-type states, and voters in y-type states benefit at the others' expense. Finally, in case 3, voters' beliefs in "states' rights" may conflict with their substantive position on other issues.

Case 1: Positive Spill-ins to x*-Type States*

Examples of case 1 are casino gambling or state-run lotteries in states where voters are concerned only with tax revenues and jobs. Taxes imposed on products sold to out-of-state consumers also fall into case 1. In these situations, A_2 and B_3 are empty sets.[11] Thus (3) becomes $N(A_3) \gtreqless N(B_2)$.

Since an x-type law leads to positive spill-ins, it is plausible to suppose that over time more and more states pass x-type laws (i.e., the set G changes). A few states institute lotteries, for example, and eventually many more follow their example. If this happens, people in set B now join set A. Many who favored a national lottery (i.e., those in B_1 and B_2) may now oppose it as a way of preserving their newly acquired positive spillovers (i.e., they move to A_3 and A_4, respectively).

As more states pass x-type laws, however, the level of net positive spillovers to each x-type state falls. In the case of casino gambling and lotteries, a state's consumers are less likely to gamble in other states if their own state has an x-type law.[12] In addition, negative spillovers to y-type states increase as more and more states export costs to them. The decline in net benefits and the increase in net costs as G changes

[11] Voters in A_2 favor preemption but oppose x in a unitary system. Voters in B_3 oppose preemption but favor policy x in a unitary system. Both of these preference patterns seem to be implausible if x-type states benefit at the expense of y-type states. This case would not hold if some people think that gambling brings costs (i.e., crowds, corruption).

[12] Similarly, tax incentives to encourage industry to locate in a jurisdiction provide few benefits to taxpayers if most states provide equivalent incentives. For an attempt to show the costs of interstate competition for business investment, see Jacobs (1979).

will raise the vote for a national preemptive law in individual x- and y-type states. Thus, a federal structure generates two influences that operate in opposite directions. A person whose state institutes a lottery is more likely to oppose federal preemption. However, as the number of states with lotteries increases, the benefits of multiplicity fall. If the first factor dominates the second, federalism will continue to matter even in the long run.

Case 2: Positive Spillovers to y-Type States

Examples of case 2 are minimum wage laws[13] or fair labor statutes that increase job opportunities in states without such laws and state pollution control laws where water and air cross state lines. Since in case 2 federalism permits y-type states to gain at the expense of x-type states, A_3 and B_2 are implausible preference patterns. Thus (3) becomes $N(B_3) \gtreqless N(A_2)$.

The majority of voters in x-type states have decided that, in spite of the costs, the law is worth having. Suppose that as time passes, voters in y-type states learn from the experience of those who have x-type (e.g., minimum wage) laws. If the benefits are greater than expected, then eventually more states pass minimum wage laws. People whose states enact such laws are now more likely to favor federal preemption. They no longer obtain the benefits of spillovers from other states, and they might like to prevent other states from benefiting at their expense. However, as the number of x-type states increases (G changes), the costs imposed on each x-type state fall and the benefits to each y-type state rise. This shift in costs and benefits should reduce support for a preemptive law in individual x- and y-type states. If the second (change in G) effect outweighs the first (change from B to A) effect, a national minimum wage law may not pass in a federal system even though many states have passed their own laws and $N(xRy) > N(yRx)$. Historical research, however, has stressed the cases where state initiatives led eventually to a federal statute,[14] leaving unanalyzed the many situations where the adoption of a law by many states has not been followed by a preemptive federal initiative.[15]

[13] Riker (1964, p. 146) uses the example of minimum wage laws that differ across states: "There is then much likelihood of capital flow from the high-wage localities to the low-wage localities for all those industries in which labor represents a high proportion of the cost. Aside from the imposition of a nationally uniform minimum wage, the only way that high-wage localities may counter this capital flow is by reducing the minimum wage level."

[14] See Fine (1956, pp. 353–85), who discusses the examples of antitrust laws, pure food and drug laws, protective labor legislation, social security, and welfare.

[15] Many of the laws studied by Walker (1969) have spread across many states but have never been adopted nationally. It is possible that some of these statutes might have majority support in a unitary system.

Case 3: States' Rights

In this case, some people put their belief in states' rights above their position on x (e.g., the abolition of slavery) and are in A_3 and B_3. Thus, no one favors imposing x on all states if they also oppose x in a unitary system (i.e., A_2 and B_2 are empty sets). Therefore, (3) becomes $N(A_3 \cup B_3) > 0$.

The vote for x in a federal system is certain to be less than the vote in a unitary system, and federalism may continue to matter so long as preferences remain constant.[16] Assume, however, that people with a substantive preference for x over y favor states' rights only if the results of independent state choices are not too far away from their position on x. Then, it is easy to tell a simple story where an unanticipated change leads to convergence of a federal and a unitary system. If the number of y-type states (e.g., slave states) increases exogenously, people who are opposed to slavery may now choose federal preemption over states' rights (i.e., some of those in A_3 and B_3 move to A_1 and B_1). Therefore, x may now pass at the national level. An exogenous change that seemed to favor y leads in fact to its repudiation by the nation. Thus, it may not be necessary to assume a growth in Northern antislavery sentiment to explain the evolution of pre–Civil War politics. As slavery spread across the South after the invention of the cotton gin, even people whose antipathy to slavery was constant might have supported a national policy of abolition if they thought that more and more states would permit slavery (i.e., they might shift from

[16] For symmetry, we could include a fourth case that appears to be of less empirical importance. In this final case, people believe in uniformity or in a strong central government and are willing to support a national law even when they would oppose the law in a unitary system. Thus, A_3 and B_3 are empty sets, (3) becomes $0 < N(A_2 \cup B_2)$, and the vote for x is always larger in a federal system than in a unitary system. If y is a status quo which permits the private market to operate with a minimum of government interference, then case 4 is illustrated by business managers who favor laissez-faire (yRx) but if faced with a mixture of differing state laws would rather have a uniform federal regulatory statute because it permits them to reduce costs. Employers who sell their products in many states often fit this preference pattern. They would rather not be regulated at all, but being regulated in some states is worse than a uniform national standard. For example, sellers of bottled mineral water are beginning to face a variety of labeling laws enacted by different states. According to *Business Week:* "Most bottlers say they would not oppose reasonable, uniform regulations. . . . Industry executives shudder at the prospect of trying to comply with rules that could vary widely among states" ("Mineral Water Could Drown in Regulation," *Business Week* [June 11, 1979]). State trucking regulations are sometimes inconsistent and costly for truckers. Thus, state regulations of the mudguards required on interstate trucks were inconsistent, and an Illinois law was overturned by the Supreme Court as interfering with interstate commerce (Bibb v. Navajo Freight Lines, Inc., 359 U.S. 520 [1959], discussed in Tribe [1978, p. 339]). Similarly, the Supreme Court ruled that Wisconsin's prohibition against twin trailers on interstate highways was an unconstitutional interference with interstate commerce ("Breaking a Bottleneck in Long-Haul Trucking," *Business Week* [March 7, 1978]).

A_3 to A_1). When slavery looked as if it might expand into the western states, people who had favored a federal solution might begin to support abolition.[17]

V. Conclusions

Political economists have generally recognized that realistic political systems must include some interjurisdictional spillovers as a cost of providing citizens with a choice of public service levels. Intergovernmental grants may reduce these interjurisdictional costs (Breton 1965) but will not eliminate them. Because both the dispersion of tastes across the population and the level of externalities differ for different public services, some analysts have argued that a federal system is the best way to accommodate these conflicting tendencies (e.g., Oates 1972). In making this recommendation, however, these authors fail to make clear an important idealization inherent in their analysis. They assume that it is possible to assign functions unambiguously to levels of governments so that constitutional structure only has an impact on spillovers and on the position of minorities. In fact, it will often be impossible to assign responsibilities neatly to a particular political level. Indeed, this overlap is the characteristic feature of contemporary "cooperative" federalism. Given this fact, the present paper has shown that, when authority is divided, the choices of lower-level governments can have important consequences for the decisions of higher-level governments. Even when the central government has the power to preempt state and local laws, its democratic choices will depend upon the strategic position of citizens living under alternative state legal regimes. The essential difference between a hierarchical federalism and a unitary system is the difference in the status quo. This difference affects the vote on national legislation and the bargaining power of individuals. Individuals may vote against the extension of a state law to the nation as a whole even though they

[17] Potter (1976), in a history of the period from 1848 to 1861, stresses the difficult trade-offs faced by many people. He writes that "the problem for Americans, who, in the age of Lincoln, wanted slaves to be free was not simply that Southerners wanted the opposite, but that they themselves cherished a conflicting value: they wanted the Constitution, which protected slavery to be honored, and the Union, which was a fellowship with slaveholders to be preserved" (pp. 44–45). The Northern public "placed their antislavery feelings in a context of state action, accepting personal responsibility for slavery within their own particular states" (p. 46). In contrasting the position of Stephen Douglas in 1854 with that of his opponents, Potter shows how these trade-offs were resolved in different people. "Douglas was a vigorous believer in the democratic principle of local autonomy, but his opponents were equally vigorous believers in the moral primacy of freedom. . . . Douglas cared more about the Union than about the eradication of slavery and would never push the slavery issue to a point where it imposed too much strain upon the Union. Many antislavery men thought the Union hardly worth preserving so long as it had slavery in it" (pp. 172–73).

would favor the law in a system with only one government. They may wish to extend a state law to all citizens although they would oppose the law in a unitary system.

The analysis suggests a promising area for future empirical research. Congressional votes on particular issues might be associated with existing state laws,[18] and the timing of federal passage of laws in several areas could be related to the patterns of state adoptions of similar laws. Empirical work is also needed on specific policy areas. For example, one might study the impact of local power over schools on state and federal education policy or see how federal efforts to reduce poverty have been conditioned by prior state and local efforts.

My discussion of federalism also suggests that, at the point of constitutional choice, people are more likely to support a federal system with strong lower-level governments, the fewer the strategic possibilities open to individual states. Strong low-level governments will be attractive if people are grouped geographically by their taste for public services and if it is difficult to impose costs on other jurisdictions. This implies that many people would support constitutional constraints on state governments that limit the states' strategic behavior while retaining many of the benefits of variety and experimentation. I would also expect that, if interjurisdictional migration is possible, then those who favor aid to currently disadvantaged groups at the expense of mobile capital and labor resources will favor a strong central government.[19] It is a commonplace in economic analyses of federalism to note that low-level governments cannot carry out redistributive policies (e.g., Oates 1972). My point, however, goes beyond this conclusion to the observation that if states try to gain at the expense of other states, then interstate redistribution can occur that bears little relation to anyone's notion of social justice and may end up making all households worse off. The end result depends upon each state's strategic position, that is, its ability to export taxes and import benefits.

References

Breton, Albert. "A Theory of Government Grants." *Canadian J. Econ. and Polit. Sci.* 31 (May 1965): 175–87.

[18] McLure (1967, p. 72), in a study of the export of state taxes, recognizes that "states with the largest export rates can be expected to favor state or local assumption of governmental activities while those with low export rates might reasonably favor federal action."

[19] Thus, Dye (1973, p. 62) writes that "urban interests, low-income groups, blacks, ethnic groups and labor organizations frequently turn to the national government for help. States' rights arguments have little appeal to those groups, which are important in the national electorate, but do not constitute majorities in the large number of sparsely settled states."

Cover, Robert M., and Aleinikoff, T. Alexander. "Dialectical Federalism: Habeas Corpus and the Court." *Yale Law J.* 86 (May 1977): 1035–1102.

Dye, Thomas R. *Politics in States and Communities.* 2d ed. Englewood Cliffs, N.J.: Prentice-Hall, 1973.

Elazar, Daniel J. *The American Partnership: Intergovernmental Co-operation in the Nineteenth-Century United States.* Chicago: Univ. Chicago Press, 1962.

Fine, Sidney. *Laissez Faire and the General-Welfare State.* Ann Arbor: Univ. Michigan Press, 1956.

Grodzins, Morton. "The Federal System." In *President's Commission on National Goals for Americans.* Englewood Cliffs, N.J.: Prentice-Hall, 1960.

———. *The American System: A New View of Government in the United States*, edited by Daniel J. Elazar. Chicago: Rand McNally, 1966.

Jacobs, Jerry. "Bidding for Business: Corporate Auctions and the 50 Disunited States." Mimeographed. Washington: Public Interest Res. Group, August 1979.

McLure, Charles E., Jr. "The Interstate Exporting of State and Local Taxes: Estimates for 1962." *National Tax J.* 20 (March 1967): 49–77.

Oates, Wallace E. *Fiscal Federalism.* New York: Harcourt Brace Jovanovich, 1972.

Olson, Mancur, Jr. "The Principle of 'Fiscal Equivalence': The Division of Responsibilities among Different Levels of Government." *A.E.R. Papers and Proc.* 59 (May 1969): 479–87.

Posner, Richard A. *Economic Analysis of Law.* 2d ed. Boston: Little, Brown, 1977.

Potter, David M. *The Impending Crisis, 1848–1861.* New York: Harper & Row, 1976.

Riker, William H. *Federalism: Origin, Operation, Significance.* Boston: Little, Brown, 1964.

———. "Federalism." In *Handbook of Political Science.* Vol. 5, *Governmental Institutions and Processes*, edited by Fred I. Greenstein and Nelson W. Polsby. Reading, Mass.: Addison-Wesley, 1975.

Rose-Ackerman, Susan. "Risktaking and Reelection: Does Federalism Promote Innovation?" *J. Legal Studies* 9 (June 1980): 593–616.

Rothenberg, Jerome. "Local Decentralization and the Theory of Optimal Government." In *The Analysis of Public Output*, edited by Julius Margolis. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1970.

Sen, Amartya K. *Collective Choice and Social Welfare.* San Francisco: Holden Day, 1970.

Tribe, Laurence H. *American Constitutional Law.* Mineola, N.Y.: Foundation Press, 1978.

Tullock, Gordon. "Federalism: Problems of Scale." *Public Choice* 6 (Spring 1969): 19–29.

Walker, Jack L. "The Diffusion of Innovations among the American States." *American Polit. Sci. Rev.* 63 (September 1969): 880–99.

———. "Innovation in State Politics." In *Politics in the American States: A Comparative Analysis*, 2d ed., edited by Herbert Jacob and Kenneth N. Vines. Boston: Little, Brown, 1971.

———. "Comment: Problems in Research on the Diffusion of Policy Innovations." *American Polit. Sci. Rev.* 67 (December 1973): 1186–91.

Wechsler, Herbert. "Political Safeguards of Federalism." In *Selected Essays on Constitutional Law 1938–1962*, edited by Edward L. Barrett, Jr., et al. St. Paul, Minn.: West Publishing (for Assoc. American Law Schools), 1963.

Wheare, Kenneth C. *Federal Government.* 3d ed. London: Oxford Univ. Press, 1953.

Winter, Ralph K., Jr. "State Law, Shareholder Protection, and the Theory of the Corporation." *J. Legal Studies* 6 (June 1977): 251–92.

Ylvisaker, Paul. "Some Criteria for a 'Proper' Areal Division of Governmental Powers." In *Area and Power: A Theory of Local Government*, edited by Arthur Maass. Glencoe, Ill.: Free Press, 1959.

Market Provision of Price-excludable Public Goods: A General Analysis

Michael E. Burns

Monash University, Melbourne

Cliff Walsh

University of Adelaide

The *demand distribution* is employed as a novel and powerful basis for analyzing monopoly provision of price-excludable public goods (and could be used to analyze other market structures or, in some circumstances, private goods provision). For (uniform) per-unit, all-or-none, two-part, and multipart pricing, we identify: characteristics of revenue functions, relative profitability, and operational procedures for selecting price-output levels. Under all these strategies, rationing of some consumers *by output* is required, and a positive price-output relationship may arise. In general the revenue ranking of pricing strategies is sensitive to the distribution of demand and, though multipart pricing can be expected to be revenue dominant, uniform per-unit pricing emerges as a surprisingly robust strategy even before operational complexity is fully taken into account.

Recent contributions to the economics literature reflect a growing interest in private provision of commodities which possess one

Work on this paper was completed while both authors were in the Department of Economics, Monash University. The notion of the demand distribution, to our knowledge, was first suggested in earlier unpublished work of the first-mentioned author. We have benefited from comments of participants in a Monash workshop, a Virginia Polytechnic Institute seminar, and the 1978 Public Choice meetings. John Head, Yew-Kwang Ng, Geoffrey Brennan, and David Friedman provided useful observations, but our greatest debt is to William Oakland, whose extensive and perceptive comments have been of considerable assistance. The authors nonetheless accept the conventional attribution of errors to themselves.

[*Journal of Political Economy*, 1981, vol. 89, no. 1]

characteristic of a public good ("jointness," or "nonrivalness in consumption") but not the other ("impossibility of exclusion"): Exclusion is assumed to be commercially feasible, so that anyone unwilling to pay the required price can be denied access to any or all of the units produced.[1] This important class of commodities can be termed *price-excludable public goods* or (reflecting a characteristic they retain rather than one they lose) *joint goods.* Television and radio transmissions and the services provided by transportation facilities (rivers, roads, airports, ships, buses, and airplanes), entertainment facilities (theaters, galleries, museums), recreational facilities (national parks, football stadiums), and information serve as examples, of varying degrees of purity:[2] All are important in modern societies, and virtually all have been suspected of being suboptimally provided by the private sector.

Competitive models dominate the published literature, and Samuelson's (1954, 1955) seminal analysis of pure public goods has clearly influenced the questions asked and the methodology employed. Analysis of the behavior of and interaction between individual profit-seeking entrepreneurs has been neglected in favor of somewhat speculative predictions of the likely characteristics of market equilibrium so that efficiency issues can be immediately examined. This is true, to some extent, even of the important contribution of Oakland (1974), in which it is suggested that competition will generate an equilibrium with unfamiliar characteristics in the joint goods context, with consumers facing an increasing step function of prices for homogeneous consumption units.[3]

In contrast, we present an analysis of joint goods provision with the individual profit-maximizing entrepreneur as its focal point. However, we depart from the analogous private goods methodology in one important respect: The familiar notion of the aggregate demand curve is replaced by the *demand distribution* as a novel, powerful, and operationally meaningful building block, reflecting the fact that the number of consumers of particular output units plays a critical role in joint goods analysis even under simple pricing strategies. Using this

[1] See, in particular, Buchanan (1967), Thompson (1968), Demsetz (1970), Oakland (1974), Auster (1977), Lee (1977), and Brennan and Walsh (1979); and also working papers by Brennan and Walsh (1978) and Burns (1979).

[2] Pure joint goods (like pure public goods) would permit an indefinitely large number of consumers. The examples cited (except perhaps information) fall well short of this but, as long as they remain uncongested, can still be analyzed as if they were "pure" (like "local" public goods). By treating costs as varying with intensity of consumption, our approach could also be applied to the impure cases.

[3] We accept Oakland's description of equilibrium, but his analysis of its attainment is essentially circular. The process of establishment of equilibrium must be substantially different from that in the private goods case. In fact, it may be that the process is analogous to the seriatim introduction of "new" goods.

basis, our central contribution is a general exposition of the neglected case of monopoly provision of joint goods, and for a variety of pricing strategies we are able to identify characteristics of revenue functions, relative profitability, and operational procedures for selecting price and output levels. Nonetheless, our framework can also provide a basis for analysis of competitive joint goods provision and for analyzing marketing strategies for *private goods,* particularly where constant marginal costs prevail.

In Section I we introduce the demand distribution, provide a geometric interpretation of it, and set out basic assumptions and definitions. Sections II and III examine the characteristics of a range of feasible marketing strategies (applied uniformly across individuals) from per-unit pricing to sophisticated multipart pricing procedures, each of which has been discussed in the private goods literature and has a basis in common business practice.[4] Section IV contains an illustrative analysis using a specific distributional form, while Section V brings together the main results with final observations.

I. Framework of the Analysis

For monopoly provision of private goods under uniform per-unit pricing, information about *aggregate* demand provides an adequate basis for decision making. Since each output unit is exclusively consumed by a single individual, the profit-maximizing price will never be less than marginal production cost, and hence the firm will always produce as many units as are demanded by all individuals at that price. The aggregate demand curve thus completely specifies the relationship between per-unit price and *both* aggregate consumption *and* production, directly determining the revenue-output relationship. However, even for a private good producer, the information revealed by the aggregate demand curve is not sufficient when discriminatory pricing schemes are being considered: for these pricing strategies, information is required on the "composition" of demand for the product.

In contrast, for monopoly production of price-excludable public goods, information on aggregate demand is inadequate even under uniform per-unit pricing. Since each production unit can be fully and equally consumed by all individuals, output need never exceed that

[4] Dupuit ([1844] 1952) was an acknowledged authority on price discrimination and influenced the adoption, by the nineteenth-century French public sector, of quite sophisticated pricing procedures. More recent literature includes general analyses by Gabor (1955–56), Puu (1964), Ng and Weisser (1974), and Yamey (1974), and contributions on quantity discounting by Buchanan (1952–53) and Berglund (1964). In many cases the analysis has been applied to goods with significant jointness elements.

required to satisfy the highest demand individual at any price. Moreover, the per-unit price faced by each individual can be less than marginal production cost since many units are jointly consumed, but this would necessitate the rationing of some high-demand individuals by output rather than by price. Consequently, not only does the conventional aggregate demand curve not define the relationship between price and output for joint goods, in general it need not even define the relationship between price and aggregate consumption. Operationally, given his knowledge of production costs, the producer of a joint good will be concerned to identify the maximum revenue obtainable from (various) given output levels, and this critically depends on the composition of demand. Specifically, he will be interested in the number of individuals who would purchase (at least) a certain quantity when confronted with a particular (revenue-maximizing) price, since this determines his marginal revenue. Compared with his private goods counterpart (who can work simply with output and price), the joint good producer must consider output, revenue-maximizing price, *and* the number of consumers of the marginal production unit. Accordingly, we have adopted as the fundamental building block of our analysis a construct which takes account of the critical distributional characteristics of demand, is based on variables that may be empirically observable in the normal process of operation, and has a basis in conventional demand theory. This construct we term the "distribution of demand" or, more succinctly, the *demand distribution.*

The Nature of the Demand Distribution

A point on the surface of the demand distribution defines the number of individuals, n, who would each consume at least q units of output if the joint good was made available at a per-unit price of p.[5] Formally (adopting the conventional private good assumptions that we are dealing with large numbers of individuals and that prices and quantities are divisible) we assume that the distribution may be represented by a well-defined function, $n = n(p,q)$, with the properties

$$\frac{\partial n}{\partial p}, \frac{\partial n}{\partial q} < 0, \tag{1}$$

so that, by the implicit function theorem, the inverse function, $p = p(n,q)$, is also well defined.

It is appropriate that we should briefly discuss the basis for these

[5] Note esp. that n does not refer to the number who would consume some part of output q, but rather to those who would consume it all.

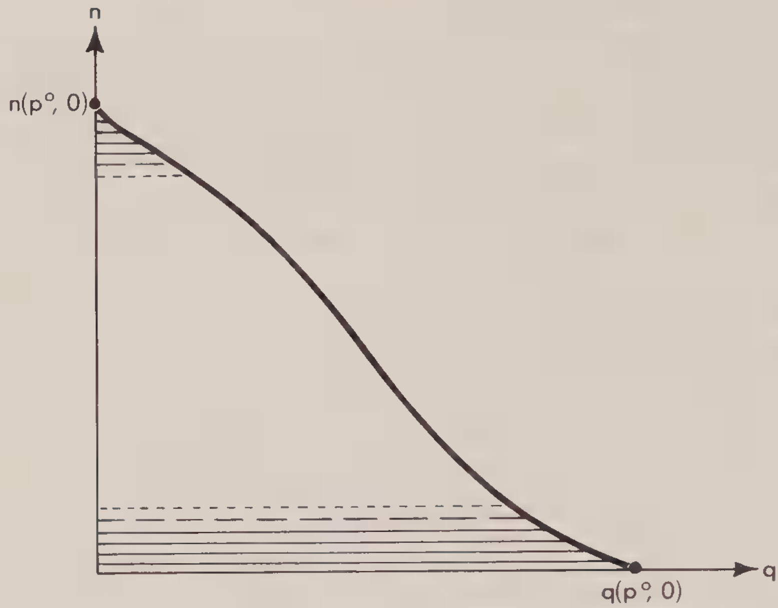


FIG. 1.—Cross section through demand distribution with price constant at p^o

specific assumptions. In figure 1 we show, for a particular per-unit price, how the number of individuals consuming each unit of output is functionally determined by the distribution of demand. For n here to be a well-defined function of q clearly we must have the demands of individuals arranged in order of magnitude; but, more than this, we require that there never be a significant number of individuals with identical demands (at any uniform price). Thus, not only have we ruled out the possibility of a significant number of identical individual demand curves but also (perhaps more interestingly), while accepting that many such curves may intersect, we have ruled out the possibility that a significant number will intersect at identical values of q (and p). The strict inequality assumed in equation (1) also involves further considerations. First, it implies that variables which influence the pattern of demand (e.g., tastes, income, and wealth) are assumed to be distributed continuously among the population. Were this not the case, significant changes in q could be associated with zero changes in n and a well-defined inverse function, $p(n,q)$, may not exist. Second, besides these implications regarding the relationship between n and q (with p constant), equation (1) also implies the assumption that individual demand curves (defined as $q_i = q_i[p]$) are downward sloping.[6]

[6] William Oakland has suggested a neat demonstration of how the distributional function, $n(p,q)$, might be derived formally from individual demand behavior. Let the inverse demand relations of a subset of individuals be given by $p = p(q,a)$, with the shift parameter distributed according to $f(a)$ over $(a,\bar{a})$. Then, for given p and q, n is determined by $n(a) = \int_a^{\bar{a}} f(s)ds$. Moreover, the intersecting case is accommodated immediately by considering any number of different classes of such demand relations. E.g., with two classes, $p = p(q,a)$ and $p = \phi(q,b)$, we would have separate density functions for a and b and define $n = n(a) + n(b)$.

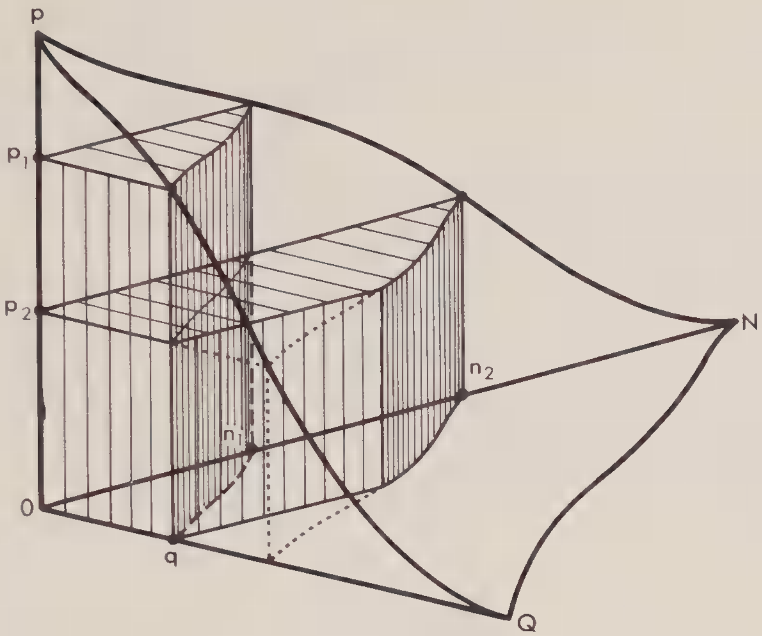


FIG. 2.—Demand distribution

The sum of all these restrictions yields the demand distribution as defined above, a geometric representation of which is provided in figure 2. Here the boundary values P, Q, and N define, respectively, the maximum price that would be paid for the initial unit of output, the maximum quantity that would be demanded by any individual at zero price, and the total number of individuals. Also shown are the solids whose volumes indicate the revenues obtained from alternative per-unit pricing arrangements associated with a particular output level. These revenue volumes reflect the fact that, for a joint good, each unit produced may be consumed by all individuals willing to pay the price, p, and are especially relevant to the analysis undertaken in Section II.

The All-or-Nothing Distribution

The incorporation into our analysis of the (desirable) feature that individual demand curves may intersect has important consequences for the treatment of all-or-nothing procedures. With respect to the perspective provided in figure 1, the presence of such intersections means that the order in which individuals' demand curves are arranged must vary according to the (constant) price level considered if ordering with respect to magnitude is to be preserved. Thus, in general it may *not* be assumed that a cross section through the distribution with n held constant will indicate the demand relation of a particular individual (although this will be so in the nonintersecting case). Even if we make the conventional assumption of zero income

effects so that $p(n,q)$ indicates the marginal evaluation of different consumption units, the integral of this inverse function, given n, cannot be assumed to measure the "willingness-to-pay" of a particular individual. For this reason, and because users of all-or-nothing procedures will generate market observations on variables different from those relevant to the per-unit pricing case, the analysis of these strategies (presented in Section III) has been undertaken with respect to an alternative (but related) distributional construct, the *all-or-nothing distribution.* A point on the surface of this distribution defines the number of individuals, n, who would purchase a bundle of q units of the good at an average (per-unit) charge C rather than go without the good altogether. We assume, analogously to the demand distribution, that the all-or-nothing distribution may be represented by a well-defined function, $n = n(C,q)$, with the properties

$$\frac{\partial n}{\partial C}, \frac{\partial n}{\partial q} < 0, \tag{2}$$

again ensuring that the inverse function, $C = C(n,q)$, is well defined.

Some Useful Definitions

Finally, later discussion is considerably assisted by reference to particular concepts which may be formally defined at this stage. The definitions made here with respect to the demand distribution may also be applied to the all-or-nothing distribution.

Definition 1: A level of output q will be termed *efficient with respect to (a particular pricing procedure)* if under that procedure: $R(q) > R(q^*)$ $\forall\ q^* < q$ (where R denotes total revenue).

Definition 2: A *regular distribution of demand* has the property: $\partial^2 R(q_0)/\partial p^2 < 0$ (where $R[q_0] = p[n,q_0] \cdot n[p,q_0]$).

Definition 3: A *weakly restricted distribution of demand* has the property that: $\partial p^*(q_0)/\partial q_0 < 0$ (where $p^*[q_0]$ is that price which maximizes $R[q_0]$).

Definitions 2 and 3 are restrictions on the nature of the demand distribution, the former requiring concavity of the revenue function relating to consumption solely in respect of the q_0th unit of output.[7] Some of the more complex operational results derived in the paper require these plausible restrictions to hold, and in these cases the

[7] Definition 2 reflects the conventional simplification adopted in private good analysis that marginal revenue functions are strictly downward sloping. The restriction implied by definition 3 is lent plausibility by the observations that it will always be true if cross sections through the distribution for q constant are (mathematically) similar and that even where this is not so the overall trend in $p^*(q)$ must be negative.

results have been presented as propositions with proofs clearly identified in the text.

II. Uniform Per-Unit Pricing

We have previously emphasized the important distinction between output and aggregate consumption in the joint goods context. Since, operationally, the entrepreneur will be concerned with the profitability of various levels of *output* and because his production cost conditions need have no unusual features, our analysis of alternative marketing strategies can focus on his search for the revenue-maximizing price (and/or charge) associated with each potential output level and, more generally, on the characteristics of the revenue functions he faces. Initially, we suppose that the joint good producer seeks to maximize profit while setting a price which is uniform across both individuals and output units.

Basic Characteristics of the Uniform Per-Unit Price

The unusual nature of his pricing decision is made transparent by first supposing that the entrepreneur behaves analogously to his private good counterpart, allowing all individuals to consume as much as they wish at the prevailing per unit price. Since, for joint goods, output would be then determined by the highest individual consumption level demanded at each price, this strategy can be alternatively represented by the assumption that at any output level the producer selects the maximum price at which anyone is willing to consume the entire output. This procedure we call *Maximum Uniform Pricing* (MUP): In figure 2, for output q it involves setting price p_1, at which price the highest demand individual consumes the entire output, and a total of n_1 individuals consume at least part of it.

The revenue yielded by this pricing procedure, for any output level q_0, is given by:

$$R(q_0)/\mathrm{MUP} = \int_0^{q_0} p_0 \cdot n(p_0,q)dq, \tag{3}$$

where $p_0 = p(0,q_0)$ and is constant throughout the integration. Differentiation of (3) with respect to q_0 yields the marginal revenue (MR) function:[8]

$$\begin{aligned} \mathrm{MR}(q_0)/\mathrm{MUP} = {} & \frac{\partial p_0}{\partial q_0} \int_0^{q_0} n(p_0,q)dp \\ & + p_0 \int_0^{q_0} \frac{\partial}{\partial p} n(p_0,q) \frac{\partial p}{\partial q} dq + p_0 \cdot n(p_0,q_0). \end{aligned} \tag{4}$$

[8] Discussion of the appropriate algebraic methods for derivation of this expression is to be found in Courant (1960), p. 220.

Note, however, that the last term in this expression, representing the effect of a change in output, price held constant, in the limit approaches zero because $n(p_0,q_0)$ itself tends to zero. Thus, $\mathrm{MR}(q_0)/\mathrm{MUP}$ can be written as:

$$\mathrm{MR}(q_0)/\mathrm{MUP} = \frac{\partial}{\partial p}[R(q_0)/\mathrm{MUP}]\frac{\partial p}{\partial q}. \tag{4a}$$

This marginal revenue function has a number of interesting properties (identified in the illustrated example in fig. 4 below), but more importantly it immediately suggests that for joint goods the optimal (i.e., revenue-maximizing) pricing strategy differs from MUP (the private goods analogue). Any level of output which is *efficient with respect to MUP* must have $\mathrm{MR}(q)/\mathrm{MUP} > 0$, and it follows directly from (4) and (4a) that, at any such output, revenue would be increased by lowering price (with output held constant). This is clearly illustrated in figure 2, where for output q, lowering price from p_1 (i.e., $p[q]/\mathrm{MUP}$) to p_2 increases total revenue: revenue declines from the original consumption units, but this is more than offset by the fact that each output unit is now more intensively consumed. Thus, *the Optimal Uniform Price, p(q)/OUP, will be less than the Maximum Uniform Price, p(q)/MUP, at any output level where marginal cost is positive.* Moreover, *some individuals* (all those for whom $q_i[p] > q$) *will be rationed by output while others* (for whom $q_i[p] \leq q$) *are rationed by price,* and this is indeed a general characteristic of optimal pricing for joint goods which does not apply for private goods.[9]

Some General Properties of Revenue Functions under OUP

Some insights into more general properties of the relevant revenue functions are straightforward. If OUP is charged for an output level q_0, we must have

$$\frac{\partial[R(q_0)/\mathrm{OUP}]}{\partial p} = 0. \tag{5}$$

It follows trivially, either from arguments similar to (3) and (4) above or from geometric considerations, that:

$$\mathrm{MR}(q_0)/\mathrm{OUP} = [p(q_0)/\mathrm{OUP}] \cdot n[p(q_0)/\mathrm{OUP},q_0]. \tag{6}$$

[9] The possible need to ration by output for per-unit pricing was first recognized in Brennan and Walsh (1978). Strictly speaking, rationing applies only if at the profit-maximizing output MC > 0 (see subsection, "Some General Properties of Revenue Functions under OUP," following). Moreover, while for $p > \mathrm{MC} > 0$ rationing cannot sensibly be considered to apply, the range of output levels for which this occurs is negligible when large numbers of consumers are involved. It should be observed, also, that the use of continuous mathematical methods makes it appropriate to regard $n[p(q_0)/\mathrm{OUP},q_0] > 0$ as indicative of rationing even for the special case of $1 > n[p(q_0)/\mathrm{OUP},q_0] > 0$.

That is, *under OUP marginal revenue (with respect to output) will be equal to the revenue obtained solely in respect of the final (i.e., least consumed) unit of output.* Regarding the characteristics of the function MR(q)/OUP, it clearly can never exceed max $[p(n,q) \cdot n(p,q)]$ which, given the properties of the demand distribution (1) above, must decrease as output increases. Since MR(0)/OUP = max $[p(n,0) \cdot n(p,0)]$ marginal revenue under OUP must obtain its maximum value at zero output, although it may sometimes increase with output at some nonzero production levels.[10]

The results (5) and (6) also yield insight into the characteristics of OUP when output is selected to yield the maximum possible revenue. Since this situation must involve $p(q)$/OUP > 0 but MR(q)/OUP $= 0$ it follows from (6) that there will be no rationing. In this case the OUP solution has MUP characteristics and, since (5) always holds for OUP, is in fact identical to the revenue-maximizing solution for MUP.

Output and Price Determination under OUP

Since profit-maximization requires the equality of marginal cost and marginal revenue, we have begun by deriving and identifying the properties of the marginal revenue function appropriate to OUP, properties which (with respect to output) turn out to be similar to those observed in the private good case. The actual process of output determination under OUP is not unduly complicated. Once the Optimal Uniform Price has been (experimentally) determined for a particular output level it follows directly from (6), providing consumption of the final unit of output is known or can be estimated, that marginal revenue is also determined. However, the decision to increase or decrease output must be accompanied by a decision regarding price adjustment, and here the user of OUP is confronted, in complete contrast to the conventional (private goods) wisdom, by the possibility that the profit-maximizing price may rise as the profit-maximizing output increases, ceteris paribus. In order to show this we first derive a general result:

PROPOSITION 1: The Optimal Uniform Price for a given output level, $p(q)$/OUP, may generally be expected to lie in the range bounded by the previous such price, $p(q - dq)$/OUP, and the price which maximizes revenue solely in respect of the qth unit of output, $p^*(q)$. This will always be so if the demand distribution is *regular.*

PROOF: If a distribution is regular, for a given q and dq, the revenue

[10] The operationality of expression (6) does, of course, require that it is possible to identify how many individuals consume the entire output. Lack of this information in no way invalidates the results derived for total and marginal revenue functions throughout the paper, but it must be recognized that practical determination of MR often may be less straightforward than is implied by the simplicity of the stated results.

expression $p(n,q) \cdot n(p,q)dq$ is a concave function of price, as therefore must be the sum of any such expressions. The revenue obtained from per-unit pricing q units of output may be expressed as $\int_0^{(q-dq)} p(n,q) \cdot n(p,q)dq + p(n,q) \cdot n(p,q)dq$, which is therefore the sum of two concave functions of price. As such its maximum must be attained at a price bounded by those prices which maximize each of its parts, $p(q - dq)$/OUP and $p^*(q)$. Even if the functions are not globally concave with respect to price, since they are continuous they must be locally concave at their maxima. Where the maxima are not at significantly different prices, as well as in those cases where convexities in the revenue functions are not severe, we might also expect $p(q)$/OUP to be bounded as suggested. □

This conclusion, which is useful in the derivation of later results, also has some operational content. It suggests that in the absence of information about the trend of $p^*(q)$, larger output expansions could initially be accompanied by a price *decrease,* and for *small* increases in output (since $p^*[q + dq] \rightarrow p^*[q]$) perhaps price initially should be set between $p(q)$/OUP and $p^*(q)$. More immediately, however, it leads to a fundamental observation:

PROPOSITION 2: The Optimal Uniform Price, $p(q)$/OUP, may sometimes increase as output increases. However, this will never occur if the demand distribution is *regular* and *weakly restricted.*

PROOF: The first part follows directly from the possibility (illustrated in fig. 3, where $p^*[1]$ applies strictly to the initial unit of output and hence is equal to $p[1]$/OUP) that $p(q - dq)$/OUP may be less than

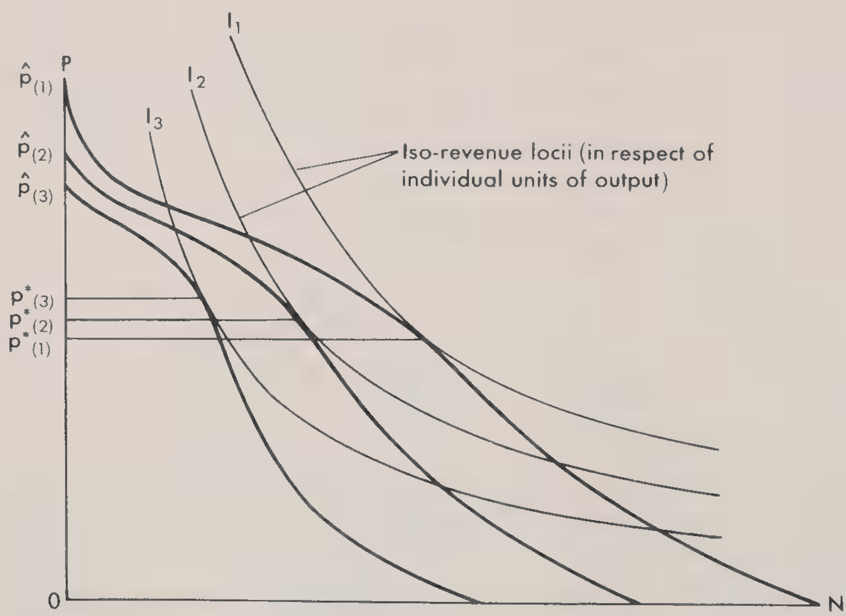


FIG. 3.—Distribution characteristics yielding price increases as output increases under OUP.

$p^*(q)$. For the second part, given a weakly restricted distribution, so that $p^*(q - dq) > p^*(q)$, and regularity, it follows from proposition 1 that

$$
\begin{aligned}
&p(0)/\mathrm{OUP} && = p^*(0) && > p^*(dq) \\
\therefore\ &p(0)/\mathrm{OUP} && > p(dq)/\mathrm{OUP} && > p^*(dq) > p^*(2dq) \\
\therefore\ &p(q - dq)/\mathrm{OUP} && > p(q)/\mathrm{OUP} && > p^*(q). \quad \square
\end{aligned}
$$

The phenomenon of (optimal) price rising when output is increased (i.e., for reductions in marginal cost) appears to be unique to the joint goods context. Moreover, even in the case of regular and weakly restricted distributions, the more conventional fall in price (as output rises) may be very small compared with that expected for a private good.[11] Other unusual phenomena, though perhaps to be given no more attention than peculiarities which can arise in traditional private goods cases, may also occur. For example, the optimal number of individuals to ration may (perversely) increase as output increases.

An Overview of Results for Uniform Per-Unit Pricing

Except in the special case where marginal cost is zero, Optimal Uniform Pricing of joint goods for any output involves a price lower than the Maximum Uniform Price for that output, and, consequently, some individuals will be rationed by output. Marginal revenue for OUP can be defined straightforwardly as the revenue obtained solely in respect of the last unit of output; the characteristics of the MR function are similar to those applying to private goods; and prescriptions regarding price and output determination can be obtained with only slightly more difficulty than in the private goods case. However, the Optimal Uniform Price may increase as output increases, and, even if this result does not arise, the price reduction associated with output increases may be small relative to those which would occur if the good was private.

III. More Complex Pricing Strategies

Observation of common business practice suggests that in the course of their marketing operations firms can generate information which

[11] Brennan and Walsh (1978) isolate a result in which price increases with output, but one which applies to specific points (actually "gaps") in the price-output relationship and which appears to emerge from the small-numbers case they deal with. Our analysis, however, using large numbers, confirms that the whole trend of prices may be upward over some output ranges. Our other result—that price falls may be small—is illustrated by the linear distribution example in Section IV since $P/2 > p(q)/\mathrm{OUP} > P/3$, where P is the price intercept of the highest demand individual.

permits them to adopt pricing strategies more sophisticated than uniform per-unit pricing. Given the "service" nature of the typical examples, monopoly producers of joint goods may be well placed to consider the use of all-or-nothing charges, entrance or license fees, and procedures which price units (or bundles of units) separately, though on operational grounds we assume that such procedures are applied uniformly across consumers. As with per-unit pricing, our analysis of these possibilities is concerned to identify basic characteristics of the optimal pricing arrangement and operational aspects of price and output determination. We are now also concerned, however, to consider the relative profitability of the alternatives open to the firm, and for this purpose the demand and all-or-nothing distributions become all the more important. However, to enable clear-cut results to be generated we are obliged to assume that income effects are negligible. Since substitution effects can generally be expected to dominate income effects, it is likely that our results are broadly applicable, but the possibility of conclusions being reversed by income effects in some instances cannot be ruled out.

All-or-Nothing Pricing

All-or-nothing type pricing schemes are exemplified by "season ticket only" offers for a series of operas or Shakespeare plays and flat-rate bus or subway fares. We assume that the same all-or-nothing offer confronts all individuals for a particular output bundle offered by the producer, though it is convenient to express this offer in terms of an average (per-unit equivalent) *charge* which must be paid for all units in the bundle. The Optimal (revenue-maximizing) All-or-Nothing Charge (OAN) will be such that *the marginal consumer pays his average valuation for the units in the bundle and the OAN strategy inherently involves rationing by output.*

It is interesting to note that, in the special case where individual demand curves do not intersect, there is an obvious symmetry between the all-or-nothing and per-unit pricing procedures; as is discussed more fully below, in that case the revenue volumes (and related algebraic expressions) may be similarly defined, except that the roles of p and n are reversed. However, to avoid these restrictions on the demand distribution we initially focus attention on the all-or-nothing distribution defined in Section I. Here, for a given output, q_0, revenue is defined by:

$$R(q_0) = q_0 \cdot C(n,q_0) \cdot n, \tag{7}$$

where C is the (per-unit) all-or-nothing charge for all units in that output bundle. The optimal value of n and associated charge C may

be obtained directly from the first-order condition:

$$\frac{\partial R(q_0)/\text{OAN}}{\partial n} = q_0\left[n \cdot \frac{\partial C}{\partial n} + C(n,q_0)\right] = 0. \tag{8}$$

This must be true for all values of q and defines n as a function of q for optimal values of C. It is straightforward to proceed from here and obtain:

$$\text{MR}(q)/\text{OAN} = n \cdot C(n,q) + \frac{\partial R(q)}{\partial n} dn^*, \tag{9}$$

where dn^* is the change in n when q is increased by one unit with C held constant (at its previous optimal level).[12]

Operationally, marginal revenue thus defined may be obtained in just two steps: First, locate the optimal per-unit charge for a given output level (which gives us $n \cdot C[n,q]$), and second, note the change in the number of consumers when output is increased by one unit and the charge (per-unit) is held constant (which yields the second part of eq. [9]). Evidently, *MR(q)/OAN is not much more difficult to determine than MR(q)/OUP.*

Of more immediate interest, however, is the revenue yield of all-or-nothing charges relative to OUP. Insight may be obtained by limiting attention to the case where individual demand curves do not intersect. In these circumstances, $R(q_0)$/OUP involves choosing p to maximize $\int_0^{q_0} p \cdot n(p,q)dq$ while $R(q_0)$/OAN involves choosing n to maximize $\int_0^{q_0} n \cdot p(n,q)dq$ (in the nonintersecting case $p[n,q]$ being, for given n, the nth individual's inverse demand function). The natural symmetry between the two procedures leads directly to two observations. First, for any distribution which is symmetrical in p and n (i.e., so that if $p = a$, $n = b$ satisfies $n = n[p,q_0]$, so must $p = b$ and $n = a$) for some scale of measurement of these variables, the revenues must be identical; second, the existence of some distribution for which $R(q)$/OAN > $R(q)$/OUP implies the existence of a distribution yielding a contrary result. Thus, *OAN may be superior or inferior to OUP in revenue terms depending upon the characteristics of the demand distribution.* Moreover, the general properties of the two strategies may be quite similar. For example, by arguments directly analogous to those used in propositions 1 and 2 it follows that *the optimal value of C may increase as output increases, but not if the all-or-nothing distribution is both regular and weakly restricted.* It is also possible that the number consuming a given bundle, $n(q)$/OAN, may sometimes increase as output increases.[13]

[12] See again Courant (1960). As an approximation, the linear form MR(q)/OAN = MC/2 may be useful in estimating whether the unit output expansion should be tried.

[13] This can be seen easily for a distribution with nonintersecting demands (but no other restrictions), since then OAN analysis can be conducted with the basic demand

Given that selection of the Optimal All-or-Nothing Charge may be only slightly more difficult than selection of the Optimal Uniform Price, there is virtually no theoretical basis for choosing between the two procedures. This should cause little surprise because the degree of complexity underlying the maximization problem faced by the entrepreneur is identical in both cases. Alternatively, insight into the ambiguity in revenue ranking may be obtained from the realization that *C(q)OAN may be* $\lesseqgtr$ *p(q)/OUP depending upon the distribution* and (compared with a uniform per-unit price equal to C) that *OAN involves the additional consumption of (lower-valued) units by some individuals while excluding all (higher-valued) consumption by others.*[14]

Two-Part Pricing

The characteristic feature of Two-Part Pricing (TPP) is the combined use of an entrance (or license) fee and a uniform per-unit price—a combination which is commonly used in the marketing of recreational and entertainment services. The essence of the TPP case can be captured in the following:

PROPOSITION 3: The introduction of a license fee in conjunction with uniform per-unit pricing may reduce revenue. However, if the demand distribution is both regular and weakly restricted and $p(q)$/OUP is used as the per-unit price, revenue is necessarily increased.

PROOF: With uniform per-unit pricing, more individuals consume the "initial" unit of output than any other unit. The marginal introduction of a license fee may be viewed as raising the price of the initial output unit without affecting the price or consumption of later units. The revenue effect will, therefore, be

$$\frac{\partial[p(n,0) \cdot n(p,0)]}{\partial p},$$

and even for a regular distribution this may be positive or negative depending upon whether the original per-unit price was below or above $p^*(0)$ (the price which maximizes revenue from the first unit alone). It follows directly from proposition 2, however, that if OUP is

distribution but with p and n reversed: Whatever holds for OUP, an analogous result holds for OAN. This similarity between OAN and OUP, moreover, suggests an analogue to MUP which might be termed "simple all-or-nothing pricing" (SAN). For output q, it involves setting a charge so that all who have positive marginal evaluations at that output are permitted to consume, the marginal consumer paying his total evaluation. $R(q)$/SAN = $R(q)$/MUP for nonintersecting demands with the distribution symmetrical in p and n.

[14] Indeed, for the linear distribution used in Section IV we have both that $C(q)$/OAN = $p(q)$/OUP and that the additional and excluded consumptions are identical. Therefore, $R(q)$/OAN = $R(q)$/OUP.

used and the distribution is both regular and weakly restricted, the revenue effect must be positive. □

That this proposition has two distinct components is quite intentional, since together they shed light on a trade-off which gives rise to the existence of an interesting possibility. From the first part of the proposition, if a given output level has a uniform price greater than $p^*(0)$, the introduction of an entrance fee will cause a greater loss of revenue from now-excluded consumers than is gained from entrance payments by the remaining consumers. From the second part of the proposition, the two effects yield a net revenue gain for a given output level for *any* uniform price *lower* than $p^*(0)$. Insight into the distinction between the two cases may be aided by regarding the first as involving a relatively high per-unit price and small number of consumers and the second as involving the reverse.[15] However, in lowering the per-unit price to make an entrance fee profitable (in an absolute sense) the departure from the uniform per-unit price $p(q)$/OUP may be sufficiently large to involve a greater loss of revenue than can ever be captured by the fee. Thus, *for some distributions it may never be profitable (relative to OUP) to use an entrance fee in conjunction with uniform per-unit pricing (i.e., TPP).*

Notwithstanding this interesting possibility, proposition 3 does suggest that under plausible conditions some form of two-part pricing will be superior to OUP, and our conceptual experiment of introducing a *marginal* entrance fee suggests there is no great difficulty in gaining *some* extra revenue. Indeed, if for any given output level the revenue-maximizing license fee and per-unit price could be determined, there would be little difficulty in locating the profit-maximizing output level, since we must have:

$$\mathrm{MR}(q)/\mathrm{TPP} = p \cdot n(p,q), \tag{10}$$

where p is the relevant per-unit price for output level q. This unsurprising result follows from arguments similar to those used to derive MR(q)/OUP above and, indeed, it is an interesting property of many marketing strategies that the marginal revenue function is defined by the revenue obtained solely in respect of the final unit of output. The reason is straightforward. The optimal setting of the parameters for any such strategy implies that marginal changes in those parameters will not change the revenue associated with any given output level. If a marginal increase in output results in only an incremental change in the price parameter applying to final production units, then revenue with respect to the original output level will be unchanged, and *any*

[15] Strictly speaking, allowance should also be made for the way n changes as p changes in the two situations.

change in overall revenue must be simply that amount obtained solely from the final output unit.[16]

Other results (related to those for OUP) follow almost immediately from (10). Except where marginal costs are zero, *TPP will involve rationing of some individuals by output; and the per-unit price component may increase with output* but it will not do so if the demand distribution is both regular and weakly restricted. In these latter circumstances it can also be established that the per-unit price element of TPP is less than $p(q)$/OUP.[17] However, even allowing for this last insight and despite the simplicity of the marginal revenue function, the entrepreneur would still need to undertake innumerable experimental variations in both the license fee and per-unit price components at each output level. *It remains unclear, therefore, whether TPP dominates OUP (and, a fortiori, OAN) in terms of overall (operational) profitability even when it does so in terms of potential revenue-raising capacity.*

Multipart Pricing

We now suppose that the producer considers confronting consumers with a schedule of prices in which different units (or bundles of units) attract a different price (though the same price schedule is applied to all consumers). This strategy might be exemplified (though perhaps somewhat imperfectly) by "quantity discount" offers for joint goods, such as are applied to a series of orchestral concerts, operas, or plays.

Obviously, the Optimal Multipart Pricing procedure (MPP) will involve pricing each unit separately at the price, $p^*(q)$, which maximizes revenue obtained solely in respect of that unit (although, as we observe later, for operational reasons the producer may offer different sized *bundles* of units with a composite charge, or average per-unit equivalent, equal to the sum of the revenue-maximizing separate prices for the units it contains). The price-output relationship is, therefore, immediately defined and, as with other strategies, *price may increase with output over some ranges.* However, in all cases the overall *trend* must involve prices decreasing with output, and for a weakly restricted distribution (this restriction directly relating to $p^*[q]$) the price-output schedule would necessarily be negatively sloped, with higher prices charged for the most intensively used ("lower level") units of output.

[16] The reader may wish to confirm why such arguments do not apply to the MR(q)/OAN function defined with respect to parameters of the all-or-nothing distribution.

[17] Given $\partial C/\partial q$, $\partial C/\partial n < 0$, the per-unit price component of TPP will be that which maximizes revenue in respect of the last $(q - q^*)$ units, q^* being the minimum consumption of any (nonexcluded) individual. From propositions 1 and 2 this price must be less than $p(q)$/OUP given the stated restrictions.

For the MPP strategy, marginal revenue is trivially defined as the maximum revenue obtainable solely in respect of the qth unit, $\{p^*(q) \cdot n[p^*(q), q]\}$, and, given the properties of the demand distribution in (1) above, MR(q)/MPP must decrease as output increases. Equally clearly, *R(q)/MPP must exceed R(q)/OUP, except in the special case where $p^*(q) = p^*(0)$ for all q, in which case the strategies are identical in all respects.*

This superiority of MPP might have been expected to extend straightforwardly to comparisons with other relatively simple strategies, such as OAN. Interestingly enough, no such exact result can be obtained unless the restriction is made that individual demand curves do not intersect. In these special circumstances, the marginal consumer of a bundle of units under OAN necessarily has the lowest marginal valuation of each and every unit in the bundle. The revenue obtained under OAN would, therefore, be equal to that obtained from a form of multipart pricing which makes the units available at the separate prices, $p_m(q)$, given by this marginal individual's demand curve. However, *except where (quite fortuitously) those prices are also the optimal $p^*(q)$ this OAN revenue must be less than that obtained under MPP.* In more general circumstances where demand curves do intersect, it can no longer be presumed that all individuals who would consume the entire bundle at the OAN composite charge would still do so if the units were priced separately at $p_m(q)$. Even though it is also true that some individuals who would not purchase the bundle at the composite charge would undertake *some* consumption at these separate prices, it remains possible that $R(q)$/OAN may exceed the revenue obtained from the multipart pricing arrangement using $p_m(q)$. Since we do not know by how much $R(q)$/MPP exceeds this latter arrangement, we cannot be certain that it in fact exceeds $R(q)$/OAN. And this uncertainty extends to comparisons between MPP and any strategy involving a lump-sum component, including TPP. It appears that *the general expectation that MPP will be the revenue-dominant strategy, while intuitively plausible, can only be established in a probabilistic sense.* Nonetheless (and notwithstanding some operational difficulties referred to below) MPP might be regarded as closest to a theoretical "best" procedure in revenue terms when interpersonal discrimination is ruled out.

Two further aspects of the MPP strategy might be noted. First, the notion of (a schedule of) prices related to "intensity of use" immediately suggests a comparison between the MPP solution and Oakland's (1974) competitive solution. Although for Oakland's case the price schedule is unambiguously positively related to output while our monopoly counterpart may characteristically involve a negative relationship between price and output, the similarities are more than merely superficial. For a given demand distribution and marginal cost schedule, the Oakland competitive output and the monopoly output

under MPP must be identical, since both solutions must be characterized by marginal cost being equated to the "total price" $(n \cdot p)$ which maximizes revenue with respect to the last unit of output and, in both cases, this will in general be associated with rationing by output.[18] (Of course for Oakland no above-normal profit exists on any unit, and consumption by inframarginal consumers is greater for his solution than under MPP because inframarginal units bear a lower price in the competitive case.)

Second, however, given that the characteristic MPP solution involves a negatively sloped price schedule, the operational status of MPP differs markedly from that of Oakland's solution. For MPP we must assume that producers can acquire information on the relevant revenue-maximizing prices, can prevent resale (or exchange) of units, and can prevent consumers from "reversing intensity" (i.e., purchasing lower-priced units without first purchasing the higher-priced inframarginal units). Resale possibilities may be of little concern in the joint good context given the service nature of most of the examples. However, preventing reversals of intensity requires either that the producer can completely meter individual purchases or that he can present consumers with a choice between bundles of units (containing 1, 2, 3, and so on, up to q units) each with a composite charge equal to the sum of the revenue-maximizing prices for the separate units it contains (a strategy we might call MPP*). However, for MPP* to be made operational, producers must be able to determine the optimal set of composite charges (and hence the relevant $p^*[q]$), and this may be a great deal more complex than isolating the $p^*(q)$ when intensity reversals do not occur. In general, firms would initially have to reduce output to zero and proceed sequentially—first determining the revenue-maximizing charge, $p^*(1)$, for one unit of output, then producing a second unit and determining the optimal composite charge for a two-unit bundle (i.e., $CC^*[2] = p^*[1] + p^*[2]$), and so on.[19] If resale possibilities or problems of overcoming intensity reversals

[18] Oakland's output is the maximal quantity, q, for which $k \cdot p_k(q) = \text{MC}$ (ranking individuals $k = 1, \ldots, K$ so that $p_1[q]$ is the highest inverse demand). If $k \cdot p_k(q) > \text{MC}$ (for some $k > 1$) when q is such that $p_1(q) = \text{MC}$, output is expanded and so on. This is equivalent to increasing output whenever the maximum revenue from the marginal production unit exceeds MC and, unless q is maximal where $p_1(q) = \text{MC}$, at least the highest demand individual is rationed by output. A referee suggested that the equality of output for MPP and competition is analogous to the private good result for perfect discrimination and perfect competition. This may be potentially misleading: MPP is not perfect discrimination, and in general the MPP (and hence competitive) output is not efficient.

[19] Though this may be a protracted experiment and fairly costly in forgone-revenue terms (particularly if fixed costs are high), if the revenue-maximizing charge declines, firms may be able to minimize their losses by allowing consumption of further units at the per-unit price $p^*(1)$ while experimenting to find the relevant charges for bundles of two or more units. Similar but more complex procedures are required if the optimal charge may increase over some output ranges.

create insurmountable problems, then the producer will be obliged to set prices such that $p(q) \leq p(q + dq)$ for all q, and the most that can be said is that for a weakly restricted distribution, the optimal separate prices are in fact uniform and therefore coincident with OUP.

Summary of Results

The more complex, but uniformly applied, pricing strategies considered here retain two basic characteristics of optimal uniform per-unit pricing: All generally involve rationing by output, and all may generate a positive price-output relationship over some output ranges. For all the strategies, moreover, the relevant marginal revenue functions can be defined with little or no more difficulty than under OUP. Optimal All-or-Nothing pricing (OAN) is operationally little more complex than OUP, but there are no strong theoretical grounds for choosing between OAN and OUP in revenue terms. Selection of the Optimal Two-Part Price (TPP) is more difficult than selection of the OUP (and OAN), and its revenue dominance is assured only for restricted (though plausible) forms of the demand distribution. Optimal Multipart Pricing (MPP) is revenue dominant over OUP, but its dominance over OAN and TPP cannot be universally guaranteed, and its position as the "theoretical best" strategy can only be established in probabilistic terms. The MPP output level is, for given cost and demand conditions, identical to Oakland's competitive output, but for operational reasons MPP may be of somewhat limited commercial applicability.

IV. Illustrative Analysis with a Specific Distributional Form

Both to illustrate the methodology used in the general analysis and to consider some of the welfare properties of monopoly provision, we now analyze the results that emerge when a specific demand distribution is used. In choosing a distributional form, account has been taken of three factors. First, there would be heuristic gains from adopting a framework which could perform an expositional role similar to that of linear demand curves in traditional demand analysis. Second, the chosen distributional form should exhibit the more fundamental characteristics of distributions likely to be met in practice. Third, the framework should not be devoid of operational content. Consideration of these factors leads more or less unequivocally to the choice of a distribution similar to that shown in figure 2 but for which the surface is linear.[20] To enable the framework to include all-or-

[20] On this, and for a further application of results obtainable from a linear demand distribution, see Burns (1979).

nothing charges we also assume that demand curves do not intersect and that income effects may be regarded as negligible.

Within this framework, the bounding values P, Q, and N completely define the distribution and may be treated as (empirically determinable) constants throughout the analysis which follows; consideration of alternative pricing procedures can be simplified considerably by drawing on the proportionality properties of similar triangles. In this respect it is convenient to define

$$\hat{p} = \frac{p}{P}, \hat{q} = \frac{q}{Q}, \text{ and } \hat{n} = \frac{n}{N}, \tag{11}$$

where p, q, and n represent specific values of price, output, and numbers of individuals. The actual derivation of the required revenue functions and pricing rules for a range of pricing strategies has been presented in some detail in Burns (1979), and the summary information contained here in table 1 and figure 4 is sufficient for present purposes.

In table 1 seven alternative strategies are defined, the associated revenue functions and pricing rules being shown for each of these strategies. The revenue functions have been presented in a form that makes self-evident the question of relative profitability (for given output levels) and illustrates certain results obtained earlier, such as the general dominance of OUP and OAN over MUP and SAN. These functions also provide a neat illustration of two further points. First, and unsurprisingly, revenue (for given output) is positively correlated with the degree of complexity of the strategy adopted. Second, broadly speaking, schemes of comparable complexity should be expected to yield similar revenues: for example, we can see that there is nothing to choose between OUP and OAN, although nonlinear distributions could be constructed (but probably not empirically determined) to favor either of these procedures. It is also apparent from the revenue functions that the returns to increased complexity of strategy diminish very rapidly, particularly at lower output levels.

Figure 4 is fundamental in illustrating the basic methodology suggested in this paper. Entrepreneurial decision making concerning output and price determination relies heavily on information about the marginal revenue function of a chosen strategy and, of course, on the relationship of such a function to (marginal) cost behavior: figure 4 contains the derived marginal revenue functions for each strategy. Also shown is a schedule indicating the sum of marginal valuations at different output levels: the intersection of this schedule with a marginal cost function would determine the optimally efficient output. Assuming, initially, constant marginal costs, the results for a linear distribution offer a clear warning that, over a substantial range of

TABLE 1

ALTERNATIVE PRICING STRATEGIES: DEFINITIONS AND REVENUE FUNCTIONS FOR A LINEAR DISTRIBUTION

Definition	Total Revenue	Marginal Revenue	Pricing Rule
Maximum Uniform Pricing—MUP: charging the maximum uniform price (uniform across output and individuals) such that some individual will wish to consume the entire output	$\hat{q}(2-\hat{q})^2PQN/16 - \hat{q}(2-3\hat{q})^2PQN/16$	$\hat{q}(2-3\hat{q})PN/2$	$p = (1-\hat{q})P$
Simple All-or-Nothing Charging—SAN: selecting the all-or-nothing charge for a bundle of q units such that all individuals who have a nonnegative evaluation of the qth unit shall be willing to pay that charge	...	...	$C = \hat{q}P/2$
Optimal Uniform Pricing—OUP: for a given output level q, charging the uniform price that maximizes revenue	$\hat{q}(2-\hat{q})^2PQN/16$	$(\hat{q}-1)^2PN/4 - \hat{q}^2PN/16$	$p = (2-\hat{q})P/4$
Optimal All-or-Nothing Charging—OAN: for a given output level q, selecting the all-or-nothing charge that maximizes revenue	...	...	$C = (2-\hat{q})P/4$
Optimal Two-Part Pricing—TPP: for a given output choosing the combination of lump-sum charge plus a uniform price that maximizes revenue	$\hat{q}(2-\hat{q})^2PQN/16 + \hat{q}^3PQN/64$	$(\hat{q}-1)^2PN/4 - \hat{q}^2PN/64$	Entrance fee $\hat{q}P/4$ $p = (4-3\hat{q})P/8$
Separate Pricing of Bundles—SPB: jointly choosing the optimal partition of output into two bundles and an optimal all-or-nothing charge for each bundle	...	...	Optimal bundles are equal size: $p_1 = (4-\hat{q})P/8$ $p_2 = (4-3\hat{q})P/8$
Optimal Multipart Pricing—MPP (or MPP*): assuming that each unit of output (or each consumption level—MPP*) may be priced (charged) separately, selecting those prices (charges) for each unit (bundle) that maximize revenue	$\hat{q}(2-\hat{q})^2PQN/16 + \hat{q}^3PQN/48$	$(\hat{q}-1)^2PN/4$	$p_i = (1-\hat{q}_i)P/2$

NOTE.—Both total and marginal revenue functions are defined for $\hat{q} = q/Q$. The algebraic expressions are based upon those derived in Burns (1979) and have been presented in comparable rather than simplified forms.

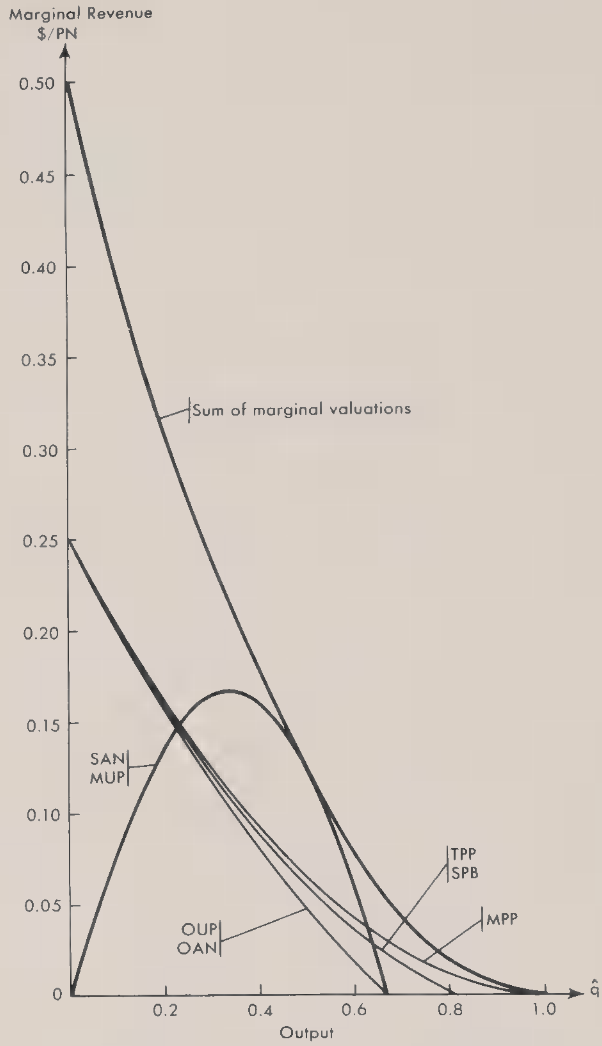


FIG. 4.—Marginal revenue functions for a linear distribution

(higher) marginal cost values, there may be negligible differences in the outputs achieved by different procedures of varying complexity and that the outputs attained by even the most complex procedures may be considerably less than the optimally efficient output. On the other hand, at lower levels of marginal cost the difference in output levels becomes more pronounced, indicative of the fact that more complex procedures achieve their maximum possible revenues at higher output levels, but in general all procedures more nearly approach the optimally efficient output level, especially the more sophisticated ones. Finally, noting that the monotonicity of MR(q)/MPP is guaranteed while that of other marginal revenue functions is plausible,[21] a situation of decreasing marginal costs will increase the

[21] E.g., although MR(q)/OUP may sometimes increase with output, its general trend must be downward (see Section II, "Some General Properties of Revenue Functions

divergence in output levels, while increasing marginal costs would have the converse effect.

In conclusion, we offer a cautionary remark about comparisons between privately achieved and optimally efficient output levels. Our results indicate (not surprisingly) that suboptimal output levels can be expected for all of the private procedures considered. While these results are relevant to the concern shown by a number of authors who have made such comparisons in other contexts (and have sometimes claimed that efficient outputs might be secured privately), it is clear that comparisons of output levels alone are of little interest. In the case of the linear distribution, outputs under OUP and OAN would be identical, but the efficiency implications of the two procedures are significantly different. Indeed, one of the insights generated by the approach advanced in this paper is that the replacement of a uniform per-unit pricing strategy with a (comparably complex) all-or-nothing arrangement will succeed only in substituting the consumption of lower-valued units of output by higher-demand individuals for the consumption of higher-valued units of output by lower-demand individuals. In general, efficiency comparisons must be based on the degree of consumption rationing (whether by price or by output) that occurs for all consumers vis-à-vis the completely efficient outcome in which all individuals would consume all units produced and output would be such that the sum of marginal valuations equals marginal cost.

V. Concluding Remarks

A major contribution of this paper is its demonstration that the complete demand distribution provides a novel and powerful basis for analyzing private provision of price-excludable public (or joint) goods. For a monopoly producer of such goods we have identified, under surprisingly general conditions, the characteristics of the revenue functions associated with various pricing strategies, indicated the likely relative profitability of those strategies, and prescribed procedures which might aid firms in isolating profit-maximizing price and output for any chosen strategy. Moreover, the specific example we provide, based on a linear demand distribution, not only serves to illustrate our general results, but also in practice may constitute an important empirical basis for generating first approximations for output and price levels.

---

under OUP"). Moreover, given MPP as the theoretical best strategy, $MR(q)/MPP$ would be the boundary form toward which other MR functions tend (as complexity of the pricing strategy is increased), and $MR(q)/MPP$ is unambiguously negatively sloped.

Using this linear demand distribution, we have also sketched out some efficiency implications of our results, and, in this regard, another important application of our general approach is highlighted. For joint goods, overall efficiency depends on how output is "rationed" among consumers as much as on the level of output per se. A framework containing information on the distribution of demands clearly is indispensable to a comprehensive analysis of the efficiency of private (or, indeed, public) provision of joint goods. In fact, our framework could be used profitably as a basis for a general analysis of joint goods provision under competitive conditions and other market or nonmarket structures. Moreover, the demand distribution approach is equally essential to analysis of "discriminatory" pricing strategies for private goods, the necessary extension of the analysis presented here being particularly straightforward when constant marginal costs prevail.

References

Auster, Richard D. "Private Markets in Public Goods (or Qualities)." *Q.J.E.* 91 (August 1977): 419–30.

Berglund, Sture. "Toward a Formal Discussion of Quantity Discounts: A Dualistic Approach to Price Differentiation." *Swedish J. Econ.* 66 (December 1964): 274–85.

Brennan, Geoffrey, and Walsh, Cliff. "A Monopoly Model of Public Goods Provision: The Uniform Pricing Case." Seminar Paper no. 70, Monash Univ., Dept. Econ., January 1978.

———. "Market Provision of Public Goods: A Monopoly Version of the Oakland Model." *Finanzarchiv* 37, no. 3 (1979): 385–95.

Buchanan, James M. "The Theory of Monopolistic Quantity Discounts." *Rev. Econ. Studies* 20, no. 3 (1952–53): 199–208.

———. "Public Goods in Theory and Practice: A Note on the Minasian-Samuelson Discussion." *J. Law and Econ.* 10 (October 1967): 193–97.

Burns, Michael E. "Discrimination and Efficiency in the Pricing of Public Goods." Seminar Paper no. 82, Monash Univ., Dept. Econ., 1979.

Courant, Richard. *Differential and Integral Calculus.* Vol. 2. New York: Wiley, 1960.

Demsetz, Harold. "The Private Production of Public Goods." *J. Law and Econ.* 13 (October 1970): 293–306.

Dupuit, Jules. "On the Measure of the Utility of Public Works." Translated in *Internat. Econ. Papers*, no. 2 (1952), pp. 83–110 (*De l'utilitie et de sa mesure*, 1844; reprint ed. [with comments by M. de Bernardi and L. Einaudi], Turin, 1934).

Gabor, André. "A Note on Block Tariffs." *Rev. Econ. Studies* 23, no. 1 (1955–56): 32–41.

Lee, Dwight R. "Discrimination and Efficiency in the Pricing of Public Goods." *J. Law and Econ.* 20 (October 1977): 403–20.

Ng, Yew-Kwang, and Weisser, Mendel. "Optimal Pricing with a Budget Constraint: The Case of the Two-Part Tariff." *Rev. Econ. Studies* 41 (July 1974): 337–45.

Oakland, William H. "Public Goods, Perfect Competition, and Underproduction." *J.P.E.* 82, no. 5 (September/October 1974): 927–39.

Puu, Tõnu. "A Note about Second-Order Conditions for the Monopolist's Optimum at Differentiated Prices." *Swedish J. Econ.* 66 (September 1964): 208–14.

Samuelson, Paul A. "The Pure Theory of Public Expenditure." *Rev. Econ. and Statis.* 36 (November 1954): 387–89.

———. "Diagrammatic Exposition of a Theory of Public Expenditure." *Rev. Econ. and Statis.* 37 (November 1955): 350–56.

Thompson, Earl A. "The Perfectly Competitive Production of Collective Goods." *Rev. Econ. and Statis.* 50 (February 1968): 1–12.

Yamey, Basil. "Monopolistic Price Discrimination and Economic Welfare." *J. Law and Econ.* 17 (October 1974): 377–80.

Comments

Recalculating the Scientific Tariff

Bruce R. Bolnick

Harvard Institute for International Development

In part 2 of his classic 1960 paper, "The Cost of Protection and the Scientific Tariff," Harry G. Johnson analyzed the second-best[1] problem of using tariffs to promote "noneconomic" objectives commonly suggested as arguments for protection. Very simply, Johnson showed how to compute the balance between marginal benefits and marginal costs using Harberger-type welfare triangles. He illustrated the methodology for five selected noneconomic objectives: (*a*) a tariff to promote self-sufficiency and independence; (*b*) a tariff to promote diversification, industrialization, or agriculturalization; (*c*) a tariff to promote a "way of life"; (*d*) a tariff to increase military preparedness; and (*e*) a bargaining tariff.

Johnson assumed that world prices of importables are given, that domestic production of importables is positive before and after imposition of the tariff, and that cross-elasticity effects can be neglected.[2] Using linear demand and supply curves and free-trade values for market parameters, Johnson computed the marginal cost of an ad valorem tariff (t) to be precisely

$$MC = tP\epsilon + tC\eta, \tag{1}$$

where P and C are the free-trade levels of domestic production and consumption, and ϵ and η are the compensated own-price elasticities

I would like to express my great appreciation to Professor Edward Tower for invaluable discussions and suggestions.

[1] Strictly speaking, the use of tariffs for such purposes is not necessarily *second* best. It may be third best if the "noneconomic" objective implies a divergence in factor markets; or it may be first best if a trade divergence is involved (see Corden 1974, chap. 2).

[2] Units of output are assumed to be defined so that the world price is unity. This assumption, along with the others listed in the text, serves to simplify computations.

[*Journal of Political Economy*, 1981, vol. 89, no. 1]

of supply and demand. (I have dropped Johnson's subscript i, indicating parameters for good i, in order to simplify the notation.)

Given any specific tariff-related policy objective, the marginal benefits (MB) can be computed as a function of t and the parameters P, C, ϵ, and η. Johnson's methodology then involves setting the ratio MC/MB (call this Z) equal to unity and inferring the characteristics of the scientific tariff structure from the resulting expression.[3] However, the condition $Z = 1$ is neither necessary nor sufficient for determining the scientific tariff unless the prohibitive tariff (which satisfies $Z = 1$ only by accident) can somehow be ruled out of consideration as the second-best optimum. Furthermore, even excluding the prohibitive tariff from consideration, $Z = 1$ is sufficient only when it has a unique root in t, which will be true only for a special class of policy objectives (which define MB). Thus the scientific tariff can be inferred from the first-order conditions in only a limited set of circumstances—even if one adheres to the simplifying assumptions underlying Johnson's analysis.

These problems can be illustrated using Johnson's own case of "self-sufficiency," which he chose to measure by "the proportion of consumption supplied from domestic production." This gives

$$MB = \beta \frac{P}{C} (\epsilon + \eta)(1 - t\eta)^{-2}, \tag{2}$$

where β is the policy weight attached to marginal units of the objective, in numeraire units.[4] Thus

$$Z = \frac{MC}{MB} = \frac{tC}{\beta}\left(1 + \frac{C - P}{P}\frac{\eta}{\epsilon + \eta}\right)(1 - t\eta)^2, \tag{3}$$

which clearly is cubic in t and can therefore generate multiple roots[5] for the equation $Z = 1$. Note too that the prohibitive tariff (for which $C = P$) is at the level t_p satisfying $t_p = (C - P)/(C\eta + P\epsilon)$.

Figure 1 shows how MB and MC vary with t. Here $MB(\beta_1)$ is the marginal benefit line for a low value of β, in which case the scientific tariff (t_{s1} where $Z = 1$) is unambiguously preferred to the prohibitive tariff (t_p). For some policy weight β_2, $MB(\beta_2)$ will satisfy the property

[3] Specifically, from the total differential of Z one can determine the dt necessary to offset dX ($X = \epsilon, \eta, P$, or C), thereby characterizing the differences in t as the market parameters vary across goods or over time.

[4] I have introduced the parameter β into Johnson's model to make MB commensurate with MC, in order to clarify the discussion. This parameter is assumed to be constant, purely as a simplifying device.

[5] Johnson, in his (1960) article, avoided confronting this problem by using a linear approximation to (3) in his analysis. This device is permissible if and only if one can presume that the scientific tariff will be small, but that presumption cannot generally be valid since the whole point of the exercise is to *find* the level for the scientific tariff.

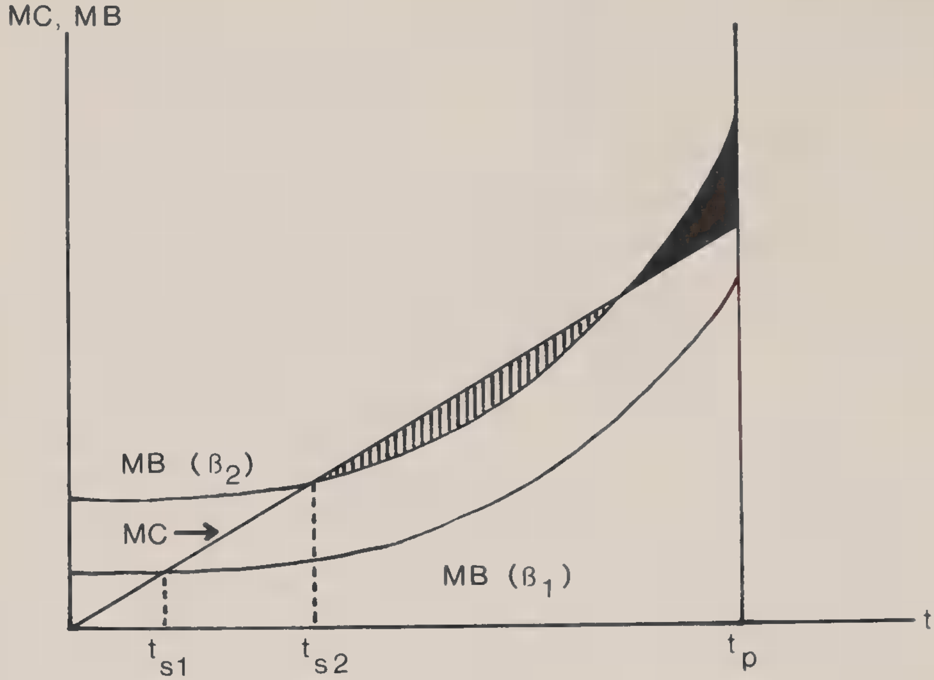


FIG. 1

that the solid area equals the slashed area, so the policymaker is indifferent between a tariff of t_{s2} (where $Z = 1$ given C, P, ϵ, and η) and one which is prohibitive ($\geqslant t_p$). For any $\beta > \beta_2$ a prohibitive tariff will always be preferred. Thus a small increase in β may cause a finite jump in the scientific tariff.[6]

Similarly, a small change in any of the market parameters may cause a finite jump in the scientific tariff (by altering MB, MC, and t_p). For example, beginning with MC and $MB(\beta_2)$ as shown in figure 1, it is clear from equation (3) that an increase in ϵ will shift MB up relative to MC, causing the scientific tariff to jump from t_{s2} to t_p (which itself is reduced in magnitude since $dt_p/d\epsilon < 0$). The possibility of such discontinuities is completely absent from Johnson's well-behaved but inappropriate formula. Furthermore, any further increase in ϵ will reduce the scientific (i.e., prohibitive) tariff, contrary to Johnson's conclusion that "the implied scientific tariff structure entails tariff rates which are higher the higher the elasticity of supply . . ." (p. 343).

In short, Johnson's methodology applies only to a limited set of circumstances; more generally, the first-order conditions are neither necessary nor sufficient for identifying the "scientific tariff."

[6] For example, if $\beta = 1$, $C = 10$, $P = 1$, and $\eta = \epsilon = 1$, then $t_p = .82$ while $t_{s2} < .33$ (since t_{s2} must be to the left of the peak of the cubic Z function, which occurs at $1/3\eta$). Thus a slight change in the planner's preference for protection, or any of the market parameters, could conceivably cause the scientific tariff to leap from below 33 percent up to 82 percent.

Moreover, the structure of protection implied by the correct calculation may differ dramatically from that implied by Johnson's procedure.

References

Corden, W. M. *Trade Policy and Economic Welfare.* Oxford: Clarendon, 1974.
Johnson, Harry G. "The Cost of Protection and the Scientific Tariff." *J.P.E.* 68, no. 4 (August 1960): 327–45.

Discount Rate and Wealth

G. S. Laumas

Illinois State University

In a recent issue of this *Journal,* Mohabbat and Simos (1978) provided estimates of the rate of discount using Kendrick's (1976) series of total private wealth. Since Kendrick also provides estimates of the division of total wealth between human, W_h, and nonhuman, W_n, wealth, it should be of interest to investigate how the discount rates differ between the two components of wealth.

In order to accomplish this, it is necessary to divide the total income into two parts—one corresponding to the human wealth, Y_h, and the other to the nonhuman wealth, Y_n. Kendrick (1976, p. 22) and Christensen and Jorgenson (1973) provide an outline of the method for such a division of the total income. Following them, the time series for Y_h and Y_n were estimated. The estimated series closely approximate the estimates for "labor compensation" and "property compensation" given by Kendrick and Christensen and Jorgenson.

As suggested in Mohabbat and Simos, the time-series estimates of the rate of discount for W_h and W_n were computed using the varying parameter regression technique. These series are given in table 1.

The obvious inference that can be drawn from the results given in table 1 is that the discount rate on W_h is consistently higher than the rate on W_n. This evidence is important by itself. In the absence of direct empirical estimates, as provided here, economists had to make assumptions about the behavior of the two rates (see Pesek and Saving 1967).

Additionally, since the major portion of W_h is education, the results provide further evidence on what Welch has called "one of the most important phenomena of our time . . . that the rates of return to investments in schooling have failed to decline under the pressure of rapidly rising average educational levels" (Welch 1970, p. 54). The existence of a relatively higher rate on W_h confirms the conclusions about the behavior of the rate of return on education arrived at by Griliches (1970, 1977), Welch (1970), Becker (1975), and others.

[*Journal of Political Economy,* 1981, vol. 89, no. 1]

TABLE I

Time-Series Estimates of the Discount Rate on Human and Nonhuman Wealth, United States, 1929 to 1969

Year	Discount Rate on Human Wealth	Discount Rate on Nonhuman Wealth
1969	.15	.08
1968	.15	.08
1967	.15	.09
1966	.15	.09
1965	.15	.09
1964	.15	.09
1963	.15	.08
1962	.16	.08
1961	.16	.08
1960	.16	.08
1959	.16	.08
1958	.16	.08
1957	.17	.08
1956	.17	.08
1955	.17	.08
1954	.17	.08
1953	.17	.08
1952	.17	.08
1951	.17	.08
1950	.17	.08
1949	.17	.08
1948	.17	.08
1947	.17	.08
1946	.18	.08
1945	.20	.08
1944	.21	.08
1943	.21	.08
1942	.20	.08
1941	.18	.07
1940	.16	.08
1939	.16	.08
1938	.16	.07
1937	.16	.07
1936	.16	.08
1935	.14	.07
1934	.14	.06
1933	.13	.06
1932	.13	.05
1931	.14	.07
1930	.15	.08
1929	.16	.09

References

Becker, Gary S. *Human Capital.* 2d ed. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1975.

Christensen, Laurits R., and Jorgenson, Dale W. "U.S. Income, Saving, and Wealth, 1929–1969." *Rev. Income and Wealth* 19 (December 1973): 329–62.

Griliches, Zvi. "Notes on the Role of Education in Production Functions and Growth Accounting." In *Education and Income*, edited by W. Lee Hansen. Studies in Income and Wealth. Vol. 35. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1970.

———. "Estimating the Returns to Schooling: Some Econometric Problems." *Econometrica* 45, no. 1 (January 1977): 1–22.

Kendrick, John W. *The Formation and Stocks of Total Capital.* New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1976.

Mohabbat, Khan A., and Simos, Evangelos E. "Consumer Horizon: Reconsidered." *J.P.E.* 86, no. 3 (June 1978): 539–41.

Pesek, Boris P., and Saving, Thomas R. *Money, Wealth, and Economic Theory.* New York: Macmillan, 1967.

Welch, Finis. "Education in Production." *J.P.E.* 78, no. 1 (January/February 1970): 35–59.

Book Reviews

Money, Inflation, and the Bank of Canada: An Analysis of Canadian Monetary Policy from 1970 to Early 1975. By Thomas J. Courchene.
Montreal: C. D. Howe Research Institute, 1976. Pp. 290. $5.00.

For 12 years, from 1950 to 1962, while other industrialized countries were operating a fixed exchange-rate regime under the Bretton Woods par value system, Canada was a maverick floater that provided economists with a laboratory in which the practical problems of implementing monetary and fiscal policy under a regime of flexible exchange rates could be observed and analyzed. Prior to the advent of greater exchange-rate flexibility among the developed countries in the early 1970s, the Canadian experience was frequently cited as an example of the degree to which, through the judicious use of monetary policy, a flexible exchange rate could be kept stable with a minimum of official intervention in the foreign exchange market. It is no coincidence that two Canadians, Robert A. Mundell and Harry G. Johnson, were major contributors to the early literature on the appropriate use of monetary and fiscal policy under alternative exchange-rate regimes.

During the late 1950s and early 1960s, debate on the relative merits of fixed versus floating exchange rates was waged within Canada not as an abstract intellectual exercise, but as a policy issue of immediate practical interest to all Canadians. Canada returned to the fold of fixed exchange-rate countries in 1962 and remained there for 8 years. When she again chose to break from the par value system on June 1, 1970 she briefly appeared once more as the odd man out. By 1974, however, the accumulating pressures of widely divergent inflation rates among countries, continuing U.S. balance of payments deficits, and rapid increases in the prices of basic commodities and energy had led to the collapse of the par value system, leaving the larger industrial countries with little choice but to opt for the present system of "managed floating" exchange rates. In this context, the experience of the Bank of Canada in conducting monetary policy under alternative exchange-rate regimes is of considerable importance to students of monetary economics and policymakers alike. It is hardly surprising that, despite its relative youth (it was established in 1934), the Bank of Canada has been the subject not only of detailed research by academic economists, but also of lively and sometimes acrimonious public debate.

Thomas J. Courchene's careful study of the Bank of Canada during the crucial and fascinating period 1970–75 is a welcome addition to a substantial literature which includes works by Neufeld (1955), Johnson and Winder

[*Journal of Political Economy*, 1981, vol. 89, no. 1]

(1962), Reuber (1964), the Royal Commission on Banking and Finance (1964), and, perhaps the best known, Paul Wonnacott's study *The Canadian Dollar, 1948–1962* (1965). Like these earlier works, Courchene's book attempts to analyze and explain the policies of the Bank of Canada within the context of standard concepts of macroeconomics and monetary theory. But Courchene's book is far from being a mere update of earlier research. His work differs in at least two ways. First, he seeks to explain his views not only to other economists but also to interested laymen, a desire which is the natural reflection of the position occupied by the Bank of Canada at center stage in the policymaking theater. Second, he approaches the analysis of monetary policy from an avowedly "monetarist" perspective. These considerations govern the structure of his work.

The book is divided into four main sections. Part 1 is devoted to an elementary elaboration of several important aspects of the monetarist theory of a small open economy under fixed and flexible exchange rates. Courchene emphasizes the distinction between nominal and real stocks of money and describes the different forces that act to equate money demand and money supply under alternative exchange-rate regimes. He then goes on to describe the concept of the natural rate of unemployment and spells out its implications for what monetary policy can and cannot do. Part 1 is completed by a brief description of alternative approaches to monetary theory, particularly the neo-Keynesian approach.

Part 2 of the book treats the theory of central banking in general and the Bank of Canada's approach in particular. Courchene first considers the various indicators of monetary stringency or ease that were used in Canada during the 1970–75 period—nominal interest rates, the liquid asset ratio of commercial banks, M1 growth, etc. He argues that the Bank of Canada's choice of policy instruments during the period was a logical consequence of its theory of how the Canadian financial system operated and particularly of the role of market imperfections in allowing the authorities to influence both "the cost of credit" and "credit conditions" to some extent independently. Throughout, Courchene has chosen the technique of comparing actual Bank of Canada policies with what would have been ideal from the point of view of monetarist theory. Such an approach is, of course, consistent with his objective, but it does involve compromises. Monetary economists of a neo-Keynesian bent will not be entirely satisfied with Courchene's rather sparse treatment of the implications of the Keynesian model for monetary policy. Courchene rightly argues that his purpose is not to create a straw man out of the neo-Keynesian approach, but it is still left to the reader to decide whether policies that do not appear optimal when viewed from a strict monetarist perspective are more reasonable from the point of view of the model that the Canadian monetary authorities had in mind.

Fortunately, this question turns out to be somewhat beside the point, for the real meat of Courchene's book is to be found in parts 3 and 4, which present a very detailed description and analysis of Canadian monetary policy during 1970–75. After an initial chapter covering the whole period, Courchene devotes a chapter to developments during each year. This method of organization, while ideal for a reference work, makes for rather heavy going when the book is being read straight through because it results in excessive repetition. Nevertheless, Courchene's achievement in these sections is impressive. He manages to present a careful and incisive analysis that integrates a great deal of statistical material on the economy's performance with the

authorities' own public explanations of their policies, drawn mainly but not exclusively from Bank of Canada publications. As a result, the comparison of actual policy with the monetarist paradigm becomes less important than the detailed description of events.

Courchene's thesis is that the floating of the Canadian dollar in June 1970 was caused by the Bank of Canada's attempt to pursue a less inflationary policy than the rest of the world, a policy that could only succeed if the exchange rate was flexible. But Courchene argues that "when the rate floated and Canada was finally allowed the independence to pursue its own policies with respect to the behavior of prices, the concern over inflation was immediately jettisoned in favor of ensuring that the exchange rate was set at an 'appropriate' level" (p. 161). This change in emphasis, in Courchene's view, ushered in a period of excessive monetary expansion that persisted from 1971 until 1975. All this is supported by the marshaling of an impressive array of facts and figures. One may choose to disagree with Courchene's views about the monetary authorities' motives and intentions, but it is not easy to dispute the broad outlines of his argument. Along the way, the reader is treated to numerous interesting details: the difficulties experienced by the authorities in using indicators and instruments based on the asset side of commercial bank balance sheets; the subtle web of problems arising out of the federal government's cash deficit and the use of Canada Savings Bond issues to finance its needs; and reliance on the transfer of government deposits, rather than standard open-market operations, to alter the level of bank reserves.

On the basis of the evidence, few would dispute Courchene's conclusion that rapid rates of monetary expansion fueled inflation in Canada in 1971–75. But this criticism is somewhat tempered when the Bank of Canada's policies are compared with the actual track record of central banks in other countries during the same period rather than with the monetarist paradigm. Furthermore, by late 1975 the Bank of Canada, like several other central banks, had announced its intention to place greater emphasis on quantitative monetary targets, a fact recognized by Courchene in a postscript apparently inserted as his book went to press. Whatever the importance of these shifts in the thrust of monetary policy, Courchene's book will undoubtedly become a standard reference work for the 1970–75 period, and the thoroughness with which he has analyzed monetary policy in Canada will serve as a firm basis for the continuing debate on this subject.

MALCOLM KNIGHT

International Monetary Fund

References

Johnson, Harry G., and Winder, John W. L. *Lags in the Effects of Monetary Policy in Canada.* Working Paper. Ottawa: Queen's Printer (for Royal Commission Banking and Finance), 1962.

Neufeld, Edward P. *Bank of Canada Operations, 1935–54.* Toronto: Univ. Toronto Press, 1955.

Reuber, Grant L. "The Objectives of Canadian Monetary Policy, 1949–61: Empirical 'Trade-Offs' and the Reaction Function of the Authorities." *J.P.E.* 72, no. 2 (April 1964): 109–32.

Royal Commission on Banking and Finance. *Report.* Ottawa: Queen's Printer, 1964.

Wonnacott, Paul. *The Canadian Dollar, 1948–1962.* Toronto: Univ. Toronto Press, 1965.

A Life for Sound Money: Per Jacobsson. His Biography. By ERIN E. JACOBSSON. Oxford: Clarendon Press, 1979. Pp. xxv + 428. £16.

Erin Jacobsson has written a lively and readable life of her father, covering the many facets of a remarkable career. She did not lack for material: 6 meters of correspondence; a diary covering over 50 years in such detail that for the later years it would fill several volumes; some 500 articles; innumerable speeches and reports, including the greater part of the Bank for International Settlements (BIS) Annual Reports for 25 years. This is drawn upon to paint a picture of the man and his many activities over the years, along with a necessarily brief account of the background to some of those activities. A concluding chapter deals with his economic thought.

From the time he joined the League of Nations Secretariat in London in the spring of 1920, until his death in office as managing director of the IMF in 1963, Per Jacobsson was almost continuously in the service of a key international organization. There was a break of 2½ years in 1929–31 when he returned to Sweden. There for a time he was economic adviser to Ivar Kreuger, the Match King, not long before the latter's suicide and the sensational collapse of Kreuger and Toll. But even in those years he spent much of his time abroad on international business. Wherever he went as a wandering economic adviser he seemed to be in the thick of the action, with an extraordinary knack of positioning himself where he could best observe, analyze, and influence events. His influence was felt over nearly half a century, not only within the international organizations for which he worked, but among those responsible for economic and financial policy throughout the world.

What was the secret of his influence as an economic adviser? Part of the answer lies in his personality. Always a little larger than life, he was the most clubbable of men, full of good stories, parables, and telling facts, persuasive and witty, confident in simple remedies and in his power to expound them, the world's outstanding salesman of sound finance. A man so entertaining and heart warming was sure of a hearing. Added to that was an extremely well-stored mind: he took endless pains to master the factual background to financial problems, to know what was going on and put together what he could learn from others about past experience and current developments. It was not for nothing that he built up an excellent daily summary of financial news at the BIS and immersed himself so deeply in the preparation of its annual reports. When he had to suggest remedies he could start from a thorough diagnosis and take full account of special circumstances without limiting himself to purely general propositions.

His approach was neither that of the economic forecaster nor that of the pure theorist. His concern was to make sure that the right policy instruments, especially monetary instruments, were brought into use and that they operated in the right direction. Thus he was no enthusiast for fine tuning and was more given to laying emphasis on keeping costs under control than on the need for demand management. He never put together any extensive analysis

of the impact of monetary policy on output and prices but took for granted that it was at once the most powerful and most delicate of instruments available for maintaining internal and external balance.

His use of economic theory was almost always at a comparatively elementary level. But he could put his finger on the key issues and had a deep intuitive understanding of the complex interactions that govern economic development. He had courage, common sense, and conviction, so that he did not hesitate to repeat unpopular doctrine until it won acceptance. Above all, he had a very good track record. He had foreseen, well in advance, the fall in long-term interest rates in the thirties. He was never much impressed by the prediction of a prolonged dollar shortage and was one of the first to stress the growth in dollar liabilities. He also realized very early—certainly by 1942—that the postwar problem was going to be one of coping with inflation rather than with unemployment. The emphasis of what he had to say moved increasingly, therefore, with the temper of the times.

The danger that he foresaw was that cheap money and deficit spending would continue in circumstances to which they were quite inappropriate. They would issue inevitably in excess demand, external deficits, and appeals for aid from abroad when policies designed to secure "internal balance" would make such aid superfluous. Such policies should be designed to restore monetary discipline, and it was on the monetary component that he kept harping until discount rates were increased in the early fifties. The message that he sought to emphasize was the primacy of monetary policy in maintaining price stability and the need simultaneously to limit public spending and preserve budgetary balance, if not year by year at least over a period of years. This was the point at which he took issue with Keynesian thinking, and it was also at the root of his popularity with central bankers.

It happens also to be the point at which he and I parted company. He had no use for the Radcliffe Committee Report of which I was a signatory. He found convincing evidence of the importance of credit control in Germany's experience in 1950–51, when an increase in the discount rate from 4 to 6 percent was followed by a gradual but pronounced swing of the German balance of payments into surplus. But this still seems to me a special case in which, for obvious reasons, stock building of imported materials had been carried to excess, and it was only a question of time before the steep upward trend in German exports predominated over speculative excesses. In general, the early postwar period was dominated by the logistical difficulties involved in regaining prewar levels of production and trade, and cheap money was not a universally important handicap in the restoration of internal balance.

However all that may be, there was no doubt about Jacobsson's priorities. He put price stability first; and since he shared neither Keynes's doubts about the self-regulating character of the economy nor Keynes's optimism about managing the level of activity with success, he was content to let employment look after itself. In some respects, therefore, he came close to the monetarist position. But he disagreed strongly with Milton Friedman on monetary management, insisting on the need for flexibility. He was also insistent on the need for what he called "a balance between prices and costs." It is not altogether clear what he meant by this, since one might suppose that the relationship between the two is governed by the state of demand. It would seem that what he was after was the danger that profit margins might be compressed by excessive wage claims or misplaced tax policies and the need to take account of cost and supply factors before deciding on the timing or scale of any boost

to demand. He envisaged the use by governments of measures designed to influence costs directly, including what has come to be called "incomes policy," so that here again he was by no means in agreement with present-day monetarists.

Although he was doubtful whether the influence of Keynes had been on balance healthy, his economic philosophy had much in common with Keynes's. He was, like Keynes, a believer in management: both the management of money and the management of the economic system through financial mechanisms. He was no believer in controls; but, like Keynes, he saw the need to find a way of stabilizing the wage level with the agreement of all parties concerned. Although he disliked deficit spending and public works, he was not prepared to condemn them in all circumstances. Even more than Keynes, he was essentially a nineteenth-century liberal, against nationalization and government subsidies, in favor of a market economy and free trade, championing the small against the large, private enterprise against government control.

The ideas for which Jacobsson contended gained increasingly in influence over the postwar world. Yet he never succeeded in committing them systematically to paper in the form of a treatise. This he regretted deeply. His compulsive need to talk—"the daily spreading of the substance of my gifts"—took precedence over the need to construct a theoretical scaffolding for his ideas. But there is no doubt that he made the right choice and that his forte lay in economic policymaking rather than in theoretical virtuosity. He left behind in his diary and his writings an ample record of the events in which he took part and of the ideas governing his approach to them.

Alec Cairncross

Oxford

Keynes' Monetary Thought: A Study of Its Development. By Don Patinkin. Durham, N.C.: Duke University Press, 1976. Pp. 155. $9.75.

It is a truly amazing story. He was a renowned teacher, scholar, and writer, an influential commentator on world affairs, an expert always consulted by his government, an intellectual active in partisan politics, a successful operator in business and finance, and much more. At the age of 47, he had just published his *magnum opus*, a two-volume treatise setting forth his mature conclusions on the entire range of his scientific expertise. The eagerly awaited work excited great interest in his profession, but considerable criticism too. The author himself was dissatisfied. Even before the ink was dry, he was changing his mind and working on the sequel. This task became a corporate enterprise, engaging not only the critical wisdom of his peers but the loyal enthusiasm of one of the most remarkable concentrations of talented young scholars in academic history. Seminars and lectures quickly turned from glosses on the completed work to draft chapters of the new. Over 5 years of constant conversation and correspondence, the master, his young disciples, and selected critics, friendly and not so friendly, argued out the new doctrines. He and his disciples never doubted they were at last figuring out how the world really works, solving puzzles that had stumped their predecessors for a century. They did not rely on experimental findings, statistical inference, historical study, or mathematical innovation, but on sheer logic and insight. Their

parochial confidence that truths worth knowing arose inside a circle of a 40-mile radius would have been ridiculous arrogance had it not been so nearly justified. Finally, the new book appeared and wrought the revolution its makers anticipated, not only commanding the attention of the profession worldwide for decades to come, but profoundly influencing the policies and politics of nations.

The two books were, of course, Keynes's *Treatise on Money* (1930) and his *General Theory* (1936). The publication of *The Collected Writings of John Maynard Keynes* by the Royal Economic Society makes available many details of the story, and more are accessible in unpublished documents among the Keynes Papers in the Marshall Library at Cambridge. It is the good fortune of the profession that a great monetary theorist, Don Patinkin, has undertaken active research in the history of monetary theory and in particular on Keynes. In the brief volume under review he traces Keynes's monetary and macroeconomic thought through his three major works, beginning with the *Tract on Monetary Reform* (1923), stressing the story sketched above of the two later books. Keynes's multiple roles in science and public life make for an engrossing case study of the interactions of theoretical developments with the events, problems, and policy controversies of the times. The intellectual and personal relationships of Keynes to his older colleagues (Pigou, Robertson, Hawtrey, Hayek), to his young disciples at Cambridge (Kahn, Joan Robinson, Austin Robinson, Sraffa) and at Oxford (Harrod, Meade), and to others (including Kaldor, Lerner, Myrdal, Ohlin, Hicks) are a fascinating episode in the sociology of scholarship. Patinkin's concise narratives whet the appetite, and he refers us to works by Winch, Moggridge, and Howson for full meals.

Here Patinkin's main emphasis is on theory. He clearly explains and compares propositions of the three books and related articles, and he seeks reasons for the changes. Most readers will be particularly interested in Patinkin's views of what Keynes "really meant" in the *General Theory.* With admirable restraint, Patinkin refrains from interpreting or appraising Keynes from the hindsight of today's theory and knowledge, even his own.

The *Tract* Patinkin finds fascinating on the national and international problems and policies of the era, but in monetary economics he finds it a routine recital of the Cambridge version of the quantity theory. The *Treatise* he finds dull and mechanical—"a Keynes out of character, a Keynes attempting to act the role of a Professor, and a Germanic one at that" (p. 24). The famous Fundamental Equations are fundamentally flawed, and the objective of explaining output fluctuations is bound to elude a theory that is designed to explain prices and profits for given output but lacks equations for output determination. It is this gap that led Keynes to the theory of effective demand, in Patinkin's view the central insight of the *General Theory.* For the *General Theory* and for Keynes's lifetime contribution Patinkin is a great enthusiast. In these days, when Keynes's standing with the profession and with the general public is cyclically low, when the man's writings and influence are blamed for all the world's ills, students and young economists will gain valuable perspective from Patinkin's book.

On the main points of interpretation I agree with Patinkin. (1) The pretension to establish an equilibrium with involuntary unemployment, which puts off so many economists instilled with neoclassical instincts, should not be taken seriously. Keynes is really describing an economy in disequilibrium, or in a succession of constrained temporary equilibria, using comparative statics as the analytical language of the day. His strong conviction and his true

message are that the automatic self-righting mechanisms of whole economies are slow, weak, and unreliable. Excess supply can persist for long periods, whether one calls the situation equilibrium or not. (2) Keynes did not assume either money wage rigidity or money illusion. He did contend that wages would adjust downward slowly—more slowly than prices could move upward. And he was skeptical of the employment effects of economy-wide wage reduction. As Patinkin says, he regarded wage and price deflation as the equivalent of monetary expansion. Neither might be capable of lowering interest rates enough to revive investment in depressed times. (3) Keynes emphasizes nonpolicy disturbances to the economy, real as well as monetary. The "state of long-term expectation" by entrepreneurs and investors is a major determinant of effective demand. Since these expectations relate to essentially unknowable and nonprobabilistic future events, including the expectations of future entrepreneurs and investors, they are arbitrary and volatile. Keynes also stresses, as in the *Treatise*, the importance of differences of view between savers and investors, or bulls and bears, or lenders and borrowers. (4) The Hicksian IS/LM apparatus captures beautifully the general equilibrium character of Keynes's vision, but its concise formalism does not do justice to the above insights and others. (In early drafts of the book and in lectures Keynes wrote symbols for expectations or "state of the news" as arguments of functions, but they were not in the book for Hicks to find.)

My differences from Patinkin's observations are minor. (1) He describes (p. 81) the financial side of the *General Theory* as a three-asset model (money, bonds, equities), in contrast to the two-asset model (money, equities) of the *Treatise*. This seems to be stretching things. It seems to me that the *General Theory* really has only two assets (money, everything else). Anyway, there is only one endogenous interest rate, that to which the estimated marginal efficiency of capital is to be equated by variation of investment—a stock-flow confusion, as Lerner observed early on. In treating the portfolio choice between equities and fixed-money-value assets and in considering the way the market brings together "bulls" and "bears," I think the *Treatise* is superior, dealing with more interesting and important matters than speculation about bond interest rates. (2) Patinkin notes with favor (p. 36) that the *Treatise* abandoned the quantity-theory presumption, followed in the *Tract*, that monetary measures first alter the quantity of money and only then, via this medium, affect spending, profits, prices, and output. In the *Treatise* central bank operations change interest rates and thereby affect all the variables, including the quantity of money, simultaneously. The *General Theory* appears to be a step backward in this respect. (3) Patinkin allows himself to wonder (p. 110) about Keynes's omission of wealth, in distinction from capital gains, as a determinant of consumption. Keynes in fact does recognize life-cycle effects, and he would properly not wish to count planned or expected accumulations of savings as another determinant. But why does he not recognize real gains or losses on money balances? The answer is not clear. Perhaps it is because none of Keynes's theoretical work contemplates drastic or continuing one-way price change. As in other monetary theory of the day, the *quaesitum* was the determination of the equilibrium value of money, that is, the price level, not the inflation rate or any other dynamic price path. Given this mind-set, Keynes might naturally regard short-run price variation as too transient and reversible to enter consumers' reckoning of wealth. The same mind-set must be the reason that Keynes did not see inflation as a way of lowering real

interest rates even when nominal rates were close to their floor. (4) Greater stress might be placed on Keynes's uncritical acceptance of the neoclassical competitive model. By assuming that firms are price takers in auction markets rather than price setters in monopolistic competition or oligopoly, he made it harder to sustain his vision of persistent disequilibrium, with failures of coordination, communication, and adjustment. Imperfect competition was the other revolution in economics in the 1930s; one of its sites was Keynes's Cambridge, and two of its agents, Joan Robinson and Sraffa, were in his group. Yet for some mysterious reason the two revolutions were never meshed.

Keynes planned, Patinkin reports, to write a book of "Footnotes to the *General Theory*," but was never able to do so. The fourth book, Patinkin feels, would have been one of continuation and clarification of the third. Keynes had no reason to rethink the *General Theory* as he had the *Treatise*.

Which of the several Keynesian traditions today is authentic? Any or all? Patinkin does not enter this dangerous terrain. We can hope he will tell us on some other occasion. Meanwhile we can look forward to more books that combine, as this one does, fascinating narrative, economic history, history of thought, and lucid exposition of difficult theory.

JAMES TOBIN
Yale University

Studies in Macroeconomic Theory. Volume 1: Employment and Inflation. By EDMUND S. PHELPS.
New York: Academic Press, 1979. Pp. xii + 418. $19.50.

This volume reprints 17 papers by Phelps that were originally published between 1965 and 1978, together with one previously unpublished paper from 1959 and some newly written material in the form of seven brief commentaries on the origins and aims of the papers and a synthetic introductory essay laying out a theory of price, wage, and employment fluctuations. Phelps's work is deep, insightful, and technically challenging. Many of the reprinted papers are well known and have been influential in the rapid development of macroeconomic analysis in recent years.

My candidate for the most significant paper in the volume is "Phillips Curves, Inflation Expectations, and Optimal Employment over Time" (1967). In this paper Phelps introduced the natural-rate hypothesis and the associated acceleration hypothesis into macroeconomic analysis, although he modestly reports that he learned these ideas from his mentors William Fellner and Henry Wallich. The natural-rate hypothesis probably has proven to be the most important new idea in macroeconomics since 1936.

Milton Friedman's apparently independent invention of the natural-rate hypothesis in his 1967 presidential address to the American Economic Association represents a remarkable coincidence in the history of the subject. Interestingly, however, Phelps and Friedman in their original presentations drew radically different lessons from the natural-rate hypothesis. For Friedman, the natural-rate hypothesis provided another reason for believing that activist stabilization policy offers little potential benefit to offset its inherent risks. For Phelps, the natural-rate hypothesis provided a new and welcomed

challenge to the technical virtuosity of students of optimal control techniques, a reaction that reflects an essential part of his implied view, discussed further below, of the relevance of economic analysis for economic policy.

The contents of this volume mostly are concerned with two problems: the development of positive models of economic behavior that can account for observed macroeconomic fluctuations, and the development of normative criteria for macroeconomic policy based on a dynamic theory of optimal inflation. With regard to the former problem, the newly written introductory essay attempts to consolidate a theory of macroeconomic fluctuations that, as Phelps puts it, is "developed piecemeal" in the reprinted papers.

In my view, this essay fails to confirm Phelps's claim that there now exists a satisfactory theory of employment fluctuations that explains why nominal disturbances apparently create a temporary divergence between actual and expected inflation rates and thereby affect real variables in addition to prices and wages. Recent years have seen progress, to which Phelps's work has contributed, in clarifying the objectives for such a theory, but have not seen the attainment of these objectives by either Phelps or other theorists. Both of the main existing approaches to this problem—the non-market-clearing models, exemplified in this volume by the introductory essay and the paper written jointly with John Taylor, "Stabilizing Powers of Monetary Policy under Rational Expectations" (1977), and the incomplete-information models, developed most fully by Robert Lucas and Robert Barro—seem to rely either explicitly or implicitly on contrived and unconvincing impediments to Walrasian outcomes.

In Phelps's new work, specifically, the key assertion is that costs of information transmission cause delays in market-clearing price and wage adjustments. This idea, however, receives neither formal theoretical development nor empirical confirmation. Phelps's substantial contributions to the ongoing work on a theory of macroeconomic fluctuations actually have largely involved qualification and embellishment of the essential clarification embodied in the natural-rate hypothesis—namely, that a Phillips curve consistent with neoclassical economic postulates would relate divergencies of unemployment from its equilibrium, that is, "natural," level to the difference between the inflation rate and some measure of the expected inflation rate, rather than simply to the inflation rate.

One accomplishment of the newly written material is to clear up ambiguities and inconsistencies in some of the reprinted papers with regard to market-clearing assumptions. As indicated above, Phelps now prefers the non-market-clearing approach, and the new introductory essay, in contrast to some of his earlier work, assumes that both labor and product markets fail to clear. His main defense of the non-market-clearing approach is that he regards the alternative incomplete-information approach as unrealistic, because when combined with assumptions about rational expectations it implies that only unsystematic components of monetary policy affect real variables. Phelps, however, discusses no empirical evidence that supports rejection of this hypothesis. Actually, the existing empirical analysis seems to be inconclusive. Moreover, here as elsewhere, his discussion of the existing literature is impressionistic and does not include either detailed examination of alternative models or even specific references.

Phelps also refers to "the patent unrealism of the view [implicit in the incomplete-information approach] that unemployment is wholly voluntary" (p. 27). However, as many authors, including myself, have pointed out, this

criticism is either semantic or wrong. Specifically, the recently developed view of labor market transactions as involving implicit contractual arrangements, which Phelps acknowledges to be a useful contribution, explains layoffs and other apparent evidence of nonwage rationing of employment without invoking a failure of labor markets to clear.

Phelps alludes in a few places to what seems to me to be the most telling and fundamental problem with the incomplete-information approach, incorporating rational expectations and market clearing, to modeling macroeconomic fluctuations. This problem is that the assumptions of this theory apparently imply that perceived monetary disturbances would not affect real variables, a hypothesis that seems inconsistent with the apparent relation between published monetary data and measures of output and employment. To get around this problem, the Lucas and Barro formulations of this theory implicitly employ the contrivance that rational individuals disregard current monetary data.

With regard to the implicitly contractual view of labor market transactions referred to above, Phelps emphasizes a variant, developed in a short paper written jointly with Guillermo Calvo, "Employment Contingent Wage Contracts" (1977), that assumes that employers know more about realized states of nature than do their workers. The significance of this assumption, however, seems questionable because, contrary to what Calvo and Phelps assume, employers in competitive labor markets probably would not find it worthwhile to exploit such an informational advantage. In any event, even with the Calvo-Phelps assumption about firm behavior, the contractual model does not imply that employment fluctuations should be related to published monetary data. Unfortunately, Calvo and Phelps obscure this point by implicitly introducing another questionable assumption, that the monetary authority has superior information about real and velocity shocks.

In addition to macroeconomic fluctuations, the other main focus of this volume is, as indicated above, a theory of optimal inflation. Phelps is interested in prescribing inflation policy both in the context of a deterministic steady state and in stochastic and non-steady-state contexts in which optimal control techniques become relevant. The analysis of these issues is comprehensive and introduces many subtleties and complexities. Examples of fundamental insights are the viewing of the problem of optimal inflation as involving the choice of an optimal tax mix and as being parallel to the problem of optimal growth. Importantly, Phelps clearly distinguishes and explains the respective roles of empirical knowledge and ethical postulates in prescribing policy.

It is worth recalling in this context that Phelps rejects models of macroeconomic fluctuations that assume market clearing and rational expectations. Such models imply that feedback control is irrelevant. In his more recent papers, however, Phelps takes careful account of how the introduction of rational expectations in place of adaptive expectations in non-market-clearing models radically changes the control problem. In "Disinflation without Recession: Adaptive Guideposts and Monetary Policy" (1978), he devises a subtle counterinflationary plan involving "*a dynamic monetary program* which is calculated to maintain full employment insofar as the average wage approximates its warranted path, and the institution of some form of *indicative wage planning* to guide and promote the growth of wages along that warranted path" (p. 256). In advocating this plan, he also carefully assesses the questions about economic behavior that make its success less than certain.

Phelps's analysis of optimal macroeconomic policy involves challenging intellectual exercise and undoubtedly enhances understanding of important economic relations. His discussions of policy issues, however, both in this volume and in other writings, also illustrate the confusion prevailing among us academic economists about the relevance of normative economic analysis for actual policymaking in real economies. In his critique, "The 1972 Report of the President's Council of Economic Advisers [CEA]: Economics and Government," published in the September 1972 issue of the *American Economic Review* (vol. 62, pp. 533–39) and referred to but not reprinted in the present volume, Phelps observes in a tone of anger and frustration "that economic analysis does not occupy a strategic place in the selection of economic policy" (p. 538). But he does not seem to be willing to entertain the possibility that the explanations for this observation may be systematic and immutable.

In criticizing the CEA, Phelps describes the so-called game plan of 1969–71 as involving both deception and self-deception, and he suggests that these symptoms are chronic characteristics of governmental inability to make effective use of scientific knowledge. In "Stopover Monetarism: Supply and Demand Factors in the 1972–74 Inflation" (1975), he attributes the record of poor monetary policy to inadequate quantitative knowledge about the economy. In his 1972 book, *Inflation Policy and Unemployment Theory* (New York: Norton), he writes of "the hope, however unwarranted, of [the cost-benefit approach to inflation planning] having an early policy impact" (p. ix). Moreover, with regard to alternative inflation policies, he acknowledges "the extraordinary difficulties of estimating costs and benefits" (p. xvii), as well as the fear that "if the lid is taken off the inflation question . . . the public will simply bungle its inflation decisions" (p. xv). At the end of this discussion, however, he succumbs to apparently wishful thinking when he writes, "I believe there could be a significant gain from adopting a less ritualistic, more 'calculating' approach to inflation policy" (p. xvi) and finally offers the advice to sink or swim when he writes, "Getting hold of our democratic policymaking potentialities is as much the point as the modest though valuable economic gain that can be expected to result" (p. xvi).

It is interesting that, in "Inflation Planning Reconsidered" (1978), Phelps discusses precautionary considerations that take account of the political inability to take appropriate action. But, he also displays unrealistic, ivory-tower faith in the importance of logic and ideas. For example, he apparently regards such constructs as "the social utility function," "the Rawlsian maximum criterion," and "the intersection of weighted individual preferences and technically feasible choices" to be potentially operational and to provide a meaningful basis for actual policymaking. In a newly written commentary, he admits to being amazed that the natural-rate hypothesis, a straightforward implication of neoclassical theorizing, turned out to be so controversial. In "Commodity-Supply Shock and Full-Employment Monetary Policy" (1978), he fancifully attributes a decision not to increase the money stock to accommodate supply shocks to Federal Reserve acceptance in 1974 of the latest, and at that date still unpublished, theoretical work implying that nonaccommodation is harmless.

Phelps clearly does not appreciate the correct perception of many laymen that much professional economic advice is based on academic fads. In the newly written material he states, "It may be that, as a methodological device, the postulate of rational expectations will prove to be irresistible—whatever

our doubts over its substantive reality" (p. 28). But does he really expect the public to take seriously prescriptions based on nothing more than professional prejudice? And how are laymen to distinguish solid advice from untested hypotheses? My guess is that attempts to apply the latest intellectual speculation to policy questions are subversive to efforts to enhance the role of fundamental economic analysis in policymaking.

Herschel I. Grossman

Brown University

Journal of Political Economy

Volume 89, Number 2, April 1981

The University of Chicago Press

JOURNAL OF

POLITICAL ECONOMY

Edited by

JACOB A. FRENKEL SAM PELTZMAN
ROBERT E. LUCAS, JR. GEORGE J. STIGLER

In cooperation with OTHER MEMBERS of the DEPARTMENT OF ECONOMICS and the GRADUATE SCHOOL OF BUSINESS of the UNIVERSITY OF CHICAGO AND OUTSIDE REFEREES

Editorial Assistants: VICKY M. LONGAWA and LISE A. PLOTKIN

The Journal of Political Economy (ISSN 0022-3808) is published bimonthly in February, April, June, August, October, and December by the University of Chicago Press. Subscription rates, U.S.A.: institutions, 1 year $30.00, 2 years $54.00, 3 years $76.50; individuals, 1 year $22.00, 2 years $39.60, 3 years $56.10. Student subscription rate, U.S.A.: 1 year $16.00 (letter from professor must accompany subscription). Other countries add $2.50 for each year's subscription to cover postage. Single copy rates: institutions $5.00, individuals $4.00. Back issues are available from 1962 (vol. 70). Make all remittances payable to *Journal of Political Economy*, The University of Chicago Press, in United States currency or its equivalent. **Business correspondence** should be addressed to The University of Chicago Press, 5801 Ellis Avenue, Chicago, Illinois 60637.

Claims for missing numbers should be made within the month following the regular month of publication. The publishers expect to supply missing numbers free only when losses have been sustained in transit and when the reserve stock will permit.

Letters to the editors and manuscripts should be addressed to the Editor of the *Journal of Political Economy*, 1126 East 59th Street, Chicago, Illinois 60637. **Manuscripts should be submitted in triplicate, accompanied by a $40.00 submission fee made payable to the Journal.** The proceeds from the submission fees are used to pay for refereeing services. Accepted manuscripts must be typed according to the University of Chicago *Manual of Style*. References should be typed double-spaced at the end of the article. Footnotes should be numbered in sequence and double-spaced following the references. Tables should follow the footnotes. Originals of the figures, drawn in india ink, should be submitted if the manuscript is accepted. Abstracts not exceeding 100 words should be submitted in duplicate along with the manuscript.

Reprinted volumes 1–72 available from Walter J. Johnson, Inc., 355 Chestnut Street, Norwood, New Jersey 07648. Volumes available **in microfilm** from University Microfilms, 300 North Zeeb Road, Ann Arbor, Michigan 48106; **in microfiche** from Johnson Associates, P.O. Box 1017, Greenwich, Connecticut 06830.

Notice to subscribers: If you change your address, please notify us and your local postmaster immediately, giving *both* your old and your new address. *Allow four weeks for the change.* **Postmaster:** Send address changes to *Journal of Political Economy*, 5801 Ellis Avenue, Chicago, Illinois 60637.

Second-class postage paid at Chicago, Illinois, and at additional mailing office.

Journal of Political Economy

Volume 89 Number 2 April 1981

Articles

Comments

Book Reviews

Interpreting Economic Time Series

Thomas J. Sargent

University of Minnesota and Federal Reserve Bank of Minneapolis

This paper explores some of the implications for econometric practice of the principle that people's observed behavior will change when their constraints change. In dynamic contexts, a proper definition of people's constraints includes among them laws of motion that describe the evolution of the taxes they must pay and the prices of the goods that they buy and sell. Changes in agents' perceptions of these laws of motion (or constraints) will in general produce changes in the schedules that describe the choices they make as a function of the information that they possess. Until very recently, received dynamic econometric practice ignored this principle. The practice of dynamic econometrics should be changed so that it is consistent with the principle that people's rules of choice are influenced by their constraints. This is a substantial undertaking and involves major adjustments in the ways that we formulate, estimate, and simulate econometric models.

Introduction

This paper explores some of the implications for econometric practice of a single principle from economic theory. This principle is that people's observed behavior will change when their constraints change. In dynamic contexts, a proper definition of people's constraints in-

The views expressed here are solely mine and do not necessarily represent the views of the Federal Reserve Bank of Minneapolis or the Federal Reserve System. Many of my thoughts on the subject of this paper have been heavily influenced by numerous discussions with Lars Peter Hansen and Robert E. Lucas, Jr. The observations on Bayesian methods are in large part those of Hansen. Ian Bain made many helpful comments on an earlier draft. This paper is the text for the Mary Elizabeth Morgan Prize lecture, given at the University of Chicago in November 1979.

[*Journal of Political Economy*, 1981, vol. 89, no. 2]

cludes among them laws of motion that describe the evolution of the taxes they must pay and the prices of the goods that they buy and sell. Changes in agents' perceptions of these laws of motion (or constraints) will in general produce changes in the schedules that describe the choices they make as a function of the information that they possess. Until very recently, received dynamic econometric practice ignored this principle and routinely deduced policy conclusions by assuming that people's rules of choice would not vary, for example, with the government's choices of laws of motion for variables such as tax rates, government purchases, and so on. These variables are supposed to have their effects precisely because they influence the constraints of some private agents.

The practice of dynamic econometrics should be changed so that it is consistent with the principle that people's rules of choice are influenced by their constraints. This is a substantial undertaking and involves major adjustments in the ways that we formulate, estimate, and simulate econometric models. Foremost, we need a stricter definition of the class of parameters that can be regarded as "structural." The body of doctrine associated with the "simultaneous equations" model in econometrics properly directs the attention of the researcher beyond reduced-form parameters to the parameters of "structural equations," which presumably describe those aspects of the behavior of people that prevail across a range of hypothetical environments. Estimates of the parameters of structural equations are needed in order to analyze an interesting class of policy interventions. Most often, however, included in a prominent way among the "structural equations" have been equations describing the rules of choice for private agents. Consumption functions, investment schedules, demand functions for assets, and agricultural supply functions are all examples of such rules of choice. In dynamic settings, regarding the parameters of these rules of choice as structural or invariant under interventions violates our simple principle from economic theory.

This paper describes methods for interpreting economic time series in a manner consistent with the principle that people's constraints influence their behavior. For the most part, I shall restrict things so that the dynamic economic theory is of the equilibrium variety, with optimizing agents and cleared markets. However, many of the principles described here will pertain to other types of dynamic economic theories, such as "disequilibrium" models with optimizing agents. The line of work I shall describe has diverse antecedents, of which major ones are contributions of Muth (1960, 1961), Nerlove (1967), Lucas and Prescott (1971), Telser and Graves (1971), and Lucas (1972*b*,

1976).[1] The works of Granger (1969) and Sims (1972) have provided key technical econometric foundations.

The basic idea is to interpret a collection of economic time series as resulting from the choices of private agents interacting in markets assumed to be organized along well-specified lines. The private agents are assumed to face nontrivial dynamic and stochastic optimization problems. This is an attractive assumption because the solutions of such problems are known to imply that the chosen variables (e.g., stocks of factors of production or financial assets) can exhibit serial correlation and cross-serial correlation. Since time series of economic data usually have the properties of high own-serial correlation and various patterns of cross-serial correlation, it seems that there is potential for specifying dynamic preferences, technologies, constraints, and rules of the market game that roughly reproduce the serial correlation and cross-correlation patterns in a given collection of time series measuring market outcomes. If this can be done in such a fashion that the free parameters of preferences, technologies, and constraints are identifiable econometrically, it is then possible to interpret the collection of time series as the outcome of a well-specified dynamic, stochastic equilibrium model. This paper is intended as a nontechnical summary of some of the econometric and theoretical issues involved in interpreting data in this way.

But why should anybody want to interpret time-series data as representing the results of interactions of private agents' optimizing choices? The answer is not that this way of modeling is aesthetically pleasing, although it is, nor that modeling in this way guarantees an analysis that implies no role for government intervention, which it does not. The reason for interpreting time series in this way is practical: potentially it offers the analyst the ability to predict how agents' behavior and the random behavior of market-determined variables will each change when there are policy interventions or other changes in the environment that alter some of the agents' dynamic constraints. There is a general presumption that private agents' behavior and the random behavior of market outcomes both will change whenever agents' constraints change, as when policy interventions or other changes in the environment occur. The most that can be hoped for is that the parameters of agents' preferences and technologies will not change in the face of such changes in the environment. If the dynamic

[1] Examples of work in the general line are Holt et al. (1960), Craine (1975), Crawford (1975), Geweke (1977), Sargent (1977, 1978), Blanco (1978), Hansen and Sargent (1979, 1980*a*, 1980*b*), Taylor (1979, 1980), Huntzinger (1979), Kennan (1979), Meese (1979), and Nerlove, Grether, and Carvalho (1979). The philosophy of this work is reviewed by Lucas and Sargent (1978, 1980).

econometric model is formulated explicitly in terms of the parameters of preferences, technologies, and constraints, it will in principle be possible for the analyst to predict the effects on observed behavior of changes in the stochastic environment.

Past dynamic econometric studies should usually be regarded as having been directed at providing ways of summarizing the observed behavior of interrelated variables, without attempting to infer the objectives, opportunities, and constraints of the agents whose decisions determine those variables. Most existing studies can be viewed, at best, as having estimated parameters of agents' decision rules for setting chosen variables as functions of the information they possess. Most of the better studies of consumption, investment, asset demand, and agricultural supply functions must be interpreted as having estimated such decision rules. Dynamic economic theory implies that these decision rules cannot be expected to remain invariant in the face of policy interventions that take the form of changes in some of the constraints facing agents. This means that there is a theoretical presumption that historical econometric estimates of such decision rules will provide poor predictions about behavior in a hypothetically new environment. This was Lucas's (1976) critique of econometric policy evaluation procedures as they existed in 1973.

Some readers of Lucas (1976) have interpreted the message as a call to evaluate policies by using existing econometric models differently.[2] However, one implication of Lucas's argument, and of dynamic economic theory generally, is that the formulation, identification, and estimation of the models must each be approached in substantially new and different ways. Most existing models simply cannot be saved by simulating them a little more shrewdly.[3]

[2] The papers by Anderson (1979) and Mishkin (1979) seem at least partly motivated by this interpretation.

[3] The set of ideas I discuss in this paper has perhaps received most notoriety in the context of macroeconomic examples. In particular, substantial attention has been devoted to the sample economies of Lucas (1972*a*) and Sargent and Wallace (1975) in which those systematic nonneutralities that come from imputing persistently suboptimal expectations to agents were shown to disappear when the hypothesis of rational expectations was imposed on agents. Crudely put, certain classes of systematic monetary policies, in particular those which operate solely via deception, were rendered impotent in the Lucas and Sargent and Wallace examples. Since the publication of these papers, many papers have been published that have described setups in which the choice of systematic policy matters, even when rational expectations prevail. These papers usually invoke a source of nonneutrality not based on deception, of which there are many in standard macroeconomic theory. Papers of this class have often been interpreted as providing a defense of "pre–rational expectations" activist policies along lines that were produced by calculating optimal controls for Keynesian econometric models of the style of the late 1960s. In fact, no such defense is implied, partly because the methods by which optimal controls for government policy variables are calculated are very different in all rational expectations models from the procedures that were applied to pre–rational expectations models, but also because the ways in which

Formulating and estimating "rational expectations" models and dynamic equilibrium models of economic time series involves a variety of important conceptual and econometric issues, some of which I try to summarize in this paper. Among the issues to be treated are the following:

i) *Identification criteria.*—Prior identifying information of the Cowles Commission variety, that is, mainly exclusion restrictions, plays a much smaller role in dynamic equilibrium models. Nonlinear cross-equation restrictions implied by dynamic theory are used extensively. This shift involves important modifications of past ways of thinking about identification and estimation.

ii) *Models of error terms.*—The dynamic equilibrium modeling strategy virtually forces the researcher to think about the sources and interpretations of the error terms in the stochastic equations that he fits. The explicitly stochastic nature of the theorizing makes it difficult to "tack on" error terms after the theorizing is done, a usual procedure in the past.

iii) *The role of Granger causality.*—Granger causality turns out to be a critical concept in the formulation of dynamic economic models, as it is coincident with the condition for the appearance as an information variable in an agent's decision rule of a variable not otherwise in the agent's criterion function or constraints.

iv) *Bayesian analysis.*—Bayesian econometric techniques provide a means of mixing prior theoretical information about parameters with information from the data. Such procedures are widely used by applied time-series econometricians, although often no formal Bayesian justification is given. Dynamic economic theory provides no justification for one widely imposed class of prior restrictions which can be viewed as restrictions directly on decision rules. Instead, dynamic economic theory suggests that prior information about agents' criterion functions and constraints is what should be used in estimation. This feature of dynamic economic theory has implications for the proper implementation both of formal Bayesian procedures and of less formal procedures for constraining parameter estimates.

I shall organize my discussion around an example, namely, a

---

econometric estimates are to be constructed for rational expectations models, with or without neutralities, differ substantially from the methods applied to the Keynesian models of the 1960s. The main point of the Lucas (1972*a*) and Sargent and Wallace (1975) examples is that substituting the assumption of rational expectations for "adaptive" expectations makes a critical difference for the methods both by which we should evaluate and optimally choose government policies. That same message is present in the papers of Fischer (1977), Phelps and Taylor (1977), and Hall (1978), even if superficially the differences in some qualitative features of the optimal policies under the two assumptions on expectations may have seemed less dramatic than in Sargent and Wallace's example or Lucas's.

linear-quadratic version of Lucas and Prescott's (1971) model of investment under uncertainty. I shall use this example for discussing the econometric implications of dynamic equilibrium models. I have adopted a linear-quadratic setup because it simplifies both the theoretical and econometric discussions, while illustrating many of the salient methodological implications of dynamic decision theory. Linear-quadratic optimum problems deliver difference equations that are linear in the variables and so match up nicely with much existing dynamic econometric theory. The reader familiar with Lucas and Prescott (1971) will recognize how the example can be generalized to incorporate more general specifications for the technologies, preferences, and constraints. That increased generality would make the econometric implications harder to extract than with the present setup, without altering the basic message.[4]

Investment under Uncertainty

This paper describes a linear-quadratic version of Lucas and Prescott's model of investment and uses it as a vehicle for expositing a variety of conceptual and econometric issues. The model describes the mutual determination over time of the capital stock, output, and market price of a single industry. The model can be generalized to handle multiple factors of production at the cost of what are really only technical complications. Similarly, the model could also be generalized to incorporate a set of industries, like the corn and hog industries, with interacting dynamics. Finally, I mention that it is straightforward to modify the model to incorporate much richer dynamics by generalizing the nature of the adjustment costs.

[4] Using the methods of discounted dynamic programming (e.g., Blackwell 1965), theoretical results establishing existence and uniqueness of equilibria and various qualitative features of the equilibria can often be obtained for "weak" or "general" assumptions, such as that utility is concave, constraint sets are convex and monotone in shift variables, and so on. Lucas and Prescott (1971) and Lucas (1978) give interesting illustrations of these methods. These techniques were also used by Sargent (1980*b*) to make some general observations on interpreting time-series correlations between Tobin's q variable and the aggregate rate of investment. However, for applied work, it is necessary to be able to calculate equilibria as a function of the free parameters of preferences and constraints, and it is highly desirable if the equilibria can be calculated easily. While for general functional forms it is in principle possible to calculate equilibria of recursive competitive models using a contraction mapping, in practice such methods are presently too expensive to use in empirical work. For this reason, for empirical work it is presently necessary to choose functional forms for which equilibria can be calculated either analytically or very quickly. Linear-quadratic specifications are one of the few such choices of convenient functional forms available. (Various versions of logarithmic specification are also sometimes tractable, e.g., Merton [1971].) A valuable treatment of recursive competitive equilibrium models with general specifications of functional forms is Prescott and Mehra (1980).

I define the following variables:

y_t = output of the representative firm;
n = number of firms in the industry, assumed constant over time;
$Y_t = ny_t$ = total output of industry;
P_t = price of output;
D_{1t} = a $(p_1 \times 1)$ vector of random variables appearing in the industry demand schedule, $p_1 \geqslant 1$;
D_{2t} = a $(p - p_1) \times 1$ vector of random variables which help predict future values of the collection of variables D_{1t}, $p \geqslant p_1$;
$D_t = \begin{bmatrix} D_{1t} \\ D_{2t} \end{bmatrix}$;
w_t = rental rate on capital;
W_t = a $(q \times 1)$ vector whose first element is w_t; the remaining elements of W_t are variables that help predict future w_t's;
u_t = a random shock to demand;
ϵ_t = a random shock in the production function;
k_t = stock of capital of the representative firm; and
$K_t = nk_t$ = total capital stock in industry.

The subscript t indexes the date to which the variable corresponds.

I further define the following polynomials in the lag operator L:

$$\delta_u(L) = 1 - \sum_{j=1}^{r_u} \delta_{uj} L^j,$$

where δ_{uj} is a scalar;

$$\delta_D(L) = I_p - \sum_{j=1}^{r_D} \delta_{Dj} L^j,$$

where δ_{Dj} is a $p \times p$ matrix and I_p is the $p \times p$ identity matrix;

$$\delta_w(L) = I_q - \sum_{j=1}^{r_w} \delta_{wj} L^j,$$

where δ_{wj} is a $(q \times q)$ matrix and I_q is the $(q \times q)$ identity matrix; and

$$\delta_\epsilon(L) = 1 - \sum_{j=1}^{r_\epsilon} \delta_{\epsilon j} L^j,$$

where $\delta_{\epsilon j}$ is a scalar.[5]

The industry consists of n identical competitive firms, each of which uses a single factor of production, capital, to produce a single output.

[5] I shall impose the condition that the zeroes of $\delta_\epsilon(z)$, $\delta_u(z)$, det $\delta_D(z)$ and det $\delta_w(z)$ each exceed unity in modulus. Actually, a weaker condition would suffice, namely, that the zeroes of these polynomials each exceed $\sqrt{\beta}$ in modulus, where β is the discount factor introduced below. These conditions on the zeroes are regularity conditions that assure that the infinite series calculated in eqq. (14) and (19) converge.

Output of the representative firm y_t is governed by

$$y_t = fk_t + n^{-1}\epsilon_t, \qquad f > 0, \tag{1}$$

where k_t is the representative firm's stock of capital at t, and ϵ_t is a random error in the technology. The firm knows $\{\epsilon_t, \epsilon_{t-1}, \ldots\}$, but does not know with certainty future values of the shock ϵ_t. The error ϵ_t is known to follow the r_ϵth-order Markov process

$$\delta_\epsilon(L)\epsilon_t = V_t^\epsilon, \tag{2}$$

where V_t^ϵ is a "fundamental" white-noise error term for ϵ_t.[6] The firm is assumed to know $\delta_\epsilon(L)$ and $E(V_t^\epsilon)^2$ with certainty.

The demand curve for output is given by[7]

$$P_t = A_0 - A_1Y_t + A_2D_{1t} + u_t, \qquad A_0, A_1 > 0, \tag{3}$$

where D_{1t} is a $(p_1 \times 1)$ vector of "demand shifters," A_2 is a $(1 \times p_1)$ vector of constants, and u_t is a random shock to the demand curve. The random term u_t obeys the r_uth-order Markov process

$$\delta_u(L)u_t = V_t^u, \tag{4}$$

where V_t^u is a fundamental white noise for u_t. The $(p_1 \times 1)$ vector of demand shifters D_{1t} consists of the first p_1 rows of the $p \times 1$ vector D_t, which follows the r_Dth-order vector autoregressive process

$$\delta_D(L)D_t = V_t^D, \tag{5}$$

where V_t^D is a $(p \times 1)$ vector white noise that is fundamental for the process D_t. The representative firm is assumed to know $\delta_u(L)$, $\delta_D(L)$, A_0, A_1, A_2, and the second moments of V_t^u and V_t^D with certainty.

At time t, total output is given by

$$Y_t = ny_t = fK_t + \epsilon_t. \tag{6}$$

The representative firm's problem is to choose a contingency plan for k_{t+j} to maximize the criterion

$$E_0 \sum_{t=0}^{\infty} \beta^t \left[P_t y_t - w_t k_t - \frac{d}{2}(k_{t+1} - k_t)^2\right], \tag{7}$$

[6] An $(n \times 1)$ vector white noise v_t^x is said to be fundamental for an $(n \times 1)$ vector process x_t if the vector of one-step-ahead linear least-squares errors in predicting x_t from past x's can be written as a linear combination of n components of v_t^x.

[7] Since a simple static demand function is posited, all of the interesting dynamics of the model come from its supply side. Specifying a demand schedule with interesting dynamics would complicate the presentation but not alter the basic messages of our example. Telser and Graves (1971) analyze dynamic optimization problems in which much of the interesting dynamics come from a demand curve that is specified. Sargent (1979, chap. 16) analyzes a model of the labor market in which the dynamics are influenced by nontrivial dynamic optimization problems solved by both suppliers and demanders.

subject to k_0 given. In (7), E_t is the mathematical expectation operator, conditional on information known to the firm at time t. This information set will shortly be specified precisely. In (7), d is a positive constant. The term $(d/2)(k_{t+1} - k_t)^2$ is intended to represent the notion that there are costs internal to the firm of adjusting the capital stock and that these rise at an increasing rate with the absolute value of the change in capital. We assume that the rental on capital w_t is the first element of the $(q \times 1)$ vector random process W_t that obeys the r_wth-order vector autoregression

$$\delta_w(L)W_t = V_t^w, \tag{8}$$

where V_t^w is a $(q \times 1)$ vector white noise that is fundamental for W_t. The firm is supposed to know $\delta_w(L)$ and the second-moment matrix of V_t^w with certainty.

At time t, the firm chooses k_{t+1}, given the information that it has available at t. However, the maximization problem (7) is not yet well posed, since we have not completely spelled out the dynamic constraints with respect to which the maximization is supposed to occur. To complete the problem (7), we begin by substituting $(fk_t + n^{-1}\epsilon_t)$ for y_t, and $(A_0 - A_1 fK_t - A_1\epsilon_t + A_2 D_{1t} + u_t)$ for P_t to get

$$E_0 \sum_{t=0}^{\infty} \beta^t \left[(A_0 - A_1 fK_t - A_1\epsilon_t + A_2 D_{1t} + u_t)(fk_t + n^{-1}\epsilon_t) \right. \\ \left. - w_t k_t - \frac{d}{2}(k_{t+1} - k_t)^2 \right]. \tag{9}$$

In order that the problem of maximizing (9) with respect to a contingency plan for $\{k_{t+j}\}$ be well posed, it is necessary to attribute to the firm precise views about the laws of motion of the random variables that it cannot control, but whose values influence the best choice of its own stocks of capital. For problem (9), these uncontrollable variables about which the representative firm cares are K_t, D_{1t}, u_t, ϵ_t, and w_t. The firm cares about the present and future behavior of the variables $(K_t, D_{1t}, u_t, \epsilon_t)$ because they influence the present and future behavior of the market price through the demand relationship $P_t = A_0 - A_1 fK_t - A_1\epsilon_t + A_2 D_{1t} + u_t$. The firm cares about the evolution of the rental process w_t because it influences its costs. We have already completely described our assumptions about the firm's views of the laws of motion of D_{1t}, u_t, ϵ_t, and w_t, namely, that the firm knows the Markov laws (4), (5), (2), and (8) that govern them, and at time t knows $D_t, D_{t-1}, \ldots, u_t, u_{t-1}, \ldots, \epsilon_t, \epsilon_{t-1}, \ldots$, and $W_t, W_{t-1}, \ldots$. To complete the specification requires that we specify the firm's views about the evolution of the aggregate capital stock K_t. We assume that the representative firm believes that the aggregate capital stock evolves according to the law

$$K_{t+1} = H_0 + H_w(L)W_t + H_D(L)D_t + H_\epsilon(L)\epsilon_t + H_u(L)u_t + H_1K_t, \tag{10}$$

where H_0 and H_1 are scalars and

$$H_w(L) = \sum_{j=0}^{r_w-1} H_{wj}L^j, \qquad \text{where } H_{wj} \text{ is } (1 \times q);$$

$$H_D(L) = \sum_{j=0}^{r_D-1} H_{Dj}L^j, \qquad \text{where } H_{Dj} \text{ is } (1 \times p);$$

$$H_\epsilon(L) = \sum_{j=0}^{r_\epsilon-1} H_{\epsilon j}L^j, \qquad \text{where } H_{\epsilon j} \text{ is a scalar; and}$$

$$H_u(L) = \sum_{j=0}^{r_u-1} H_{uj}L^j, \qquad \text{where } H_{uj} \text{ is a scalar.}$$

The representative firm is assumed to know all of the parameters of the linear law of motion (10) with certainty. The reason that we have chosen the form (10) as the firm's perceived law of motion for K will shortly become apparent.

With these specifications, the maximization of (9) is now well posed. Summarizing the setup, we have that the representative firm maximizes

$$E_0 \sum_{t=0}^{\infty} \beta^t \Big[(A_0 - A_1 fK_t - A_1\epsilon_t + A_2D_{1t} + u_t)(fk_t + n^{-1}\epsilon_t) - w_tk_t - \frac{d}{2}(k_{t+1} - k_t)^2\Big], \tag{9}$$

subject to the laws of motion[8]

$$K_{t+1} = H_0 + H_w(L)W_t + H_D(L)D_t + H_\epsilon(L)\epsilon_t + H_u(L)u_t + H_1K_t, \tag{10}$$

$$\delta_w(L)W_t = V_t^w, \tag{8}$$

$$\delta_u(L)u_t = V_t^u, \tag{4}$$

$$\delta_D(L)D_t = V_t^D, \tag{5}$$

$$\delta_\epsilon(L)\epsilon_t = V_t^\epsilon, \tag{2}$$

and subject to the information set at time t,[9]

[8] It would be straightforward to modify this setup to assume that the $\{W, u, \epsilon, D\}$ processes are each finite order mixed moving average, autoregressive processes. For the details, see Hansen and Sargent (1979).

[9] These variables completely characterize the "state" vector for the firm's problem.

$$\{K_t, k_t, W_t, W_{t-1}, \ldots, W_{t-r_w+1}, D_t, D_{t-1}, \ldots, D_{t-r_D+1}, \\ \epsilon_t, \epsilon_{t-1}, \ldots, \epsilon_{t-r_\epsilon+1}, u_t, u_{t-1}, \ldots, u_{t-r_u+1}\}.$$

The firm maximizes (9), taking the laws of motion (8), (4), (5), (2), and (10) as given and beyond its control. The firm is assumed to behave competitively and to act as if it has no control over the aggregate capital stock K. This is a reasonable assumption if n is large. The firm is assumed to know the δ's and H's with certainty and to know the first and second moments of the V_t's.[10] We further restrict the problem so that the solution is a linear contingency plan.[11] For this to be true, it is sufficient that the least-squares predictors of future W, D, ϵ, and u's be linear in the conditioning variables. This will be true if V_t^ϵ, V_t^u, V_t^D, and V_t^w obey normal probability laws. Alternatively, the analyst can simply assume that the industry is operating under optimal linear rules. In either case, the solution of the representative firm's problem is a linear contingency plan of the form[12]

$$\begin{aligned} k_{t+1} = h_0 + h_w(L)W_t + h_D(L)D_t + h_\epsilon(L)\epsilon_t \\ + h_u(L)u_t + h_1K_t + h_2k_t, \end{aligned} \tag{11}$$

where h_0, h_1, and h_2 are scalars, and

$$h_w(L) = \sum_{j=0}^{r_w-1} h_{wj}L^j, \qquad \text{where } h_{wj} \text{ is } (1 \times q);$$

$$h_D(L) = \sum_{j=0}^{r_D-1} h_{Dj}L^j, \qquad \text{where } h_{Dj} \text{ is } (1 \times p);$$

$$h_\epsilon(L) = \sum_{j=0}^{r_\epsilon-1} h_{\epsilon j}L^j, \qquad \text{where } h_{\epsilon j} \text{ is a scalar; and}$$

$$h_u(L) = \sum_{j=0}^{r_u-1} h_{uj}L^j, \qquad \text{where } h_{uj} \text{ is a scalar.}$$

The h's of (11) are in general functions both of the parameters in the criterion function (9), the parameters of δ_w, δ_u, δ_D, and δ_ϵ appearing in

---

We have in mind that the firm actually has observations on values of K, k, W, D, ϵ, and u for all dates t and earlier. It turns out that the firm's decisions are optimally a function only of the information set listed in the text.

[10] It is assumed that each of V_t^w, V_t^u, V_t^D, and V_t^ϵ is orthogonal to the information set $\{W_{t-s}, u_{t-s}, D_{t-s}, \epsilon_{t-s}, s \geq 1\}$.

[11] This is because we want the stochastic difference equations describing the behavior of the system to be linear and thereby to be readily susceptible to econometric analysis.

[12] That the solution to the problem is of this form follows from linear optimal control theory (see Kushner 1971, chap. 9; Bertsekas 1976, chap. 3; Sargent 1979, chap. 14; or Kwakernaak and Sivan 1972).

(8), (4), (5), and (2), and the H's of the perceived law of motion for capital (10). The mapping giving the h's as functions of these other parameters is defined implicitly by standard formulas in linear optimal control theory, as exposited, for example, by Kwakernaak and Sivan (1972) and Bertsekas (1976). For present purposes, it is enough to note the existence of this mapping without exploring its nature in detail. The economic content of the mapping from the δ's, H's, and objective function parameters to the h parameters of the firm's decision rule is easy to understand, since it captures the notion that the firm's rule of choice depends on both its objective and its perceived constraints (10), (8), (4), (5), and (2).

Multiplying both sides of the firm's decision rule (11) by n and using $K_t = nk_t$ gives

$$\begin{aligned} K_{t+1} = nh_0 + nh_w(L)W_t + nh_D(L)D_t + nh_\epsilon(L)\epsilon_t \\ + nh_u(L)u_t + (nh_1 + h_2)K_t. \end{aligned} \tag{12}$$

Equation (12) is the actual law of motion for aggregate capital that results from the behavior of the representative firm. The representative firm's optimization problem in effect induces a mapping from the firm's perceived law of motion for aggregate capital (10) to the actual law of motion (12). For each possible particular perceived law of motion of the form (10), there is an implied law of motion for aggregate capital of the form (12). The notion of rational expectations is that the representative firm's perceptions of (10) are correct. In effect, a rational expectations equilibrium is a fixed point of the mapping that the representative firm's optimization problem induces from (10) to (12). Formally, we define a *rational expectations equilibrium* as a perceived law of motion (10) and an implied actual law of motion (12) which are identically equal. In a rational expectations equilibrium, firms' perceptions about the law of motion for aggregate capital turn out to be confirmed by the aggregate of the choices made by firms. Upon comparing (10) with (12) it is evident that necessary and sufficient conditions for a rational expectations equilibrium are

$$\begin{aligned} H_0 &= nh_0, \\ H_w(L) &= nh_w(L), \\ H_D(L) &= nh_D(L), \\ H_\epsilon(L) &= nh_\epsilon(L), \\ H_u(L) &= nh_u(L), \\ H_1 &= (nh_1 + h_2). \end{aligned}$$

Implicit in the above definition of a rational expectations equilibrium

are the following elements: (*a*) market clearing, (*b*) optimization of the firm's expected present value, and (*c*) correct perceptions on the part of firms of the laws of motion of variables affecting their present value but beyond their control.

We begin our analysis of the model by briefly describing aspects of the optimization problem solved by the firm. Among the first-order necessary conditions for the maximization of (9) is the following system of stochastic "Euler equations," which are derived by differentiating (9) with respect to k_t for $t = 1, 2, \ldots$:

$$\begin{aligned} \beta d k_{t+1} - d(1+\beta)k_t + dk_{t-1} &= \beta w_t \\ &- \beta f(A_0 - A_1 f K_t - A_1 \epsilon_t + A_2 D_{1t} + u_t), \end{aligned} \tag{13}$$

or

$$k_{t+1} - \left(\frac{1}{\beta} + 1\right)k_t + \frac{1}{\beta}k_{t-1} = \frac{1}{d}w_t - \frac{f}{d}P_t.$$

In addition to the system of Euler equations, a transversality condition is among the first-order necessary conditions. The transversality condition can be derived by methods described in Sargent (1979). The transversality condition for the present problem in effect requires that the solution possess the property

$$\lim_{j\to\infty} E_t \beta^{t+j} k_{t+j} = 0.$$

Using the lag operator, the preceding Euler equation can be rewritten as[13]

$$\left[1 - \left(\frac{1}{\beta} + 1\right)L + \frac{1}{\beta}L^2\right]k_{t+1} = \frac{1}{d}w_t - \frac{f}{d}P_t.$$

Using the factorization

$$\left[1 - \left(\frac{1}{\beta} + 1\right)L + \frac{1}{\beta}L^2\right] = \left(1 - \frac{1}{\beta}L\right)(1 - L),$$

the above Euler equation can be written as

$$\left(1 - \frac{1}{\beta}L\right)(1 - L)k_{t+1} = \frac{1}{d}w_t - \frac{f}{d}P_t.$$

Noting that $[1 - (1/\beta)L] = -\beta^{-1}L(1 - \beta L^{-1})$ and operating on both sides of the above equation with $[-\beta^{-1}L(1 - \beta L^{-1})]^{-1}$ gives the solution[14]

[13] For a discussion of the use of lag operators in the present context, see Sargent (1979, chaps. 9 and 14).

[14] In effect, the transversality condition compels us to solve the unstable root forward in this manner.

$$(1-L)k_{t+1} = \frac{-d^{-1}\beta L^{-1}}{1-\beta L^{-1}} w_t + \frac{\beta f d^{-1} L^{-1}}{1-\beta L^{-1}} P_t,$$

or, equivalently,

$$(1-L)k_{t+1} = -d^{-1}\beta \sum_{i=0}^{\infty} \beta^i w_{t+i+1} + \beta f d^{-1} \sum_{i=0}^{\infty} \beta^i P_{t+i+1}. \tag{14}$$

It can be verified that (14) satisfies both the Euler equations and the transversality condition. Equation (14) would give the appropriate rule for setting k_{t+1} if the firm had perfect foresight about the entire future paths of the rental w_t and the output price P_t. When the firm does not have perfect foresight, the correct decision rule can be derived by replacing the future values on the right side of (14) with the corresponding mathematical expectations conditional on information the firm does have. This leads to the decision rule[15]

$$(1-L)k_{t+1} = -d^{-1}\beta \sum_{i=0}^{\infty} \beta^i E w_{t+i+1} \mid \Omega_t + \beta f d^{-1} \sum_{i=0}^{\infty} \beta^i E P_{t+i+1} \mid \Omega_t. \tag{15}$$

Here Ω_t is defined as the information set $\Omega_t = \{W_t, W_{t-1}, \ldots, u_t, u_{t-1}, \ldots, D_t, D_{t-1}, \ldots, \epsilon_t, \epsilon_{t-1}, \ldots, K_t\}$. The conditional mathematical expectations are assumed to be computed using the laws of motion (10), (8), (4), (5), and (2) for K, W, u, D, and ϵ, respectively, as well as the demand relationship $P_t = A_0 - A_1(fK_t + \epsilon_t) + A_2 D_{1t} + u_t$, which is used to deduce the law of motion for P_t. Once these conditional mathematical expectations are explicitly calculated in terms of the parameters of (10), (8), (4), (5), (2), and the demand curve (3), they can be substituted into equation (15) to deduce the optimum decision rule (11) for the representative firm. The decision rule (11) is linear in all of the information variables that appear on the right side. However, as the above method of calculating the parameters h of the decision rule (11) suggests, the parameters h are themselves complicated nonlinear functions of the underlying parameters of the model: the parameters A_0, A_1, A_2 of the demand curve, the parameters f and d of the technology, and the parameters $\delta_u(L)$, $\delta_\epsilon(L)$, $\delta_w(L)$, and $\delta_D(L)$ of the laws of motion of the random processes given from outside the model.[16] The h's are also nonlinear functions of the H's of the law of

[15] As noted above, we shall want the relevant conditional expectations to be linear. So we shall regard the $E(\cdot \mid \Omega_t)$ that appears in (15) and elsewhere as wide-sense conditional expectations, that is, linear least-squares predictors. This amounts to restricting the firm to linear decision rules, as desired.

[16] The parameters β and n also belong in this list of underlying parameters of the model. I shall usually delete these two parameters from subsequent listings of the model's underlying parameters, though they should be understood. In some applications, the analyst may want to specify counterparts of β and n completely a priori, in which case they would not be included among the free parameters of the model over which the likelihood function or other measure of "fit" is to be maximized.

motion of aggregate capital (10), which are not given from outside but are to be determined from the analysis. The nature of these nonlinearities has been characterized by Hansen and Sargent (1980*b*) and will be alluded to further below.

Equation (15), which was derived by purely formal manipulations, has the virtue of indicating clearly that the firm has an incentive to forecast future realizations of the rental w and the output price P. As a result, any state variables that the firm sees at t, and that help predict either future P's or future w's, will appear in the firm's decision rule for k_{t+1}, given by equation (11). That the h's of (11) are nonlinear functions of the parameters $\{A_0, A_1, A_2, f, d, \beta, \delta_u, \delta_\epsilon, \delta_w, \delta_D, H_u, H_\epsilon, H_w, H_D, H_0$, and $H_1\}$ stems from the nonlinear way in which the conditional mathematical expectations of future w's and P's are functions of these parameters.

In practice, to compute a rational expectations equilibrium it is not necessary ever to calculate the right side of (15). Indeed, it is never necessary explicitly to calculate the h's that determine the decision rule (11) of the representative firm. Instead, the H's of the equilibrium law of motion for the industry can be calculated directly as follows.[17] First, multiply both sides of equation (13) by n, then use $K_t = nk_t$ and collect all terms in K on the left side to get

$$\begin{aligned}\beta dK_{t+1} - [d(1+\beta) + A_1 f^2 \beta n]K_t + dK_{t-1} = \beta n w_t \\ - \beta n f A_0 + A_1 \beta f n \epsilon_t - \beta f n A_2 D_{1t} - \beta f n u_t.\end{aligned} \tag{16}$$

It is of some interest that (16) is itself the Euler equation for the "social planning" problem of maximizing[18]

$$\begin{aligned}E_0 \sum_{t=0}^{\infty} \beta^t \Big\{\Big[A_0(fK_t + \epsilon_t) - \frac{1}{2}A_1(fK_t + \epsilon_t)^2 + (fK_t + \epsilon_t)A_2 D_{1t} \\ + (fK_t + \epsilon_t)u_t\Big] - w_t K_t - \frac{1}{2} n^{-1} d(K_{t+1} - K_t)^2\Big\},\end{aligned} \tag{17}$$

[17] The following argument in the text provides a way of discovering Lucas and Prescott's (1971) method of calculating the rational expectations equilibrium by formulating a fictitious social planning problem that reproduces the equilibrium. It is worth remarking that Kydland and Prescott (1977) describe a recursive method of calculating a linear rational expectations equilibrium that is applicable to our problem and is distinct from the Lucas-Prescott method upon which the discussion in the text is based. Kydland and Prescott's method successfully computes the equilibrium even in instances in which the Lucas-Prescott method breaks down. These instances occur, e.g., in which there is feedback from the industry-wide aggregate capital stock K to W or D, as would occur if lagged K's appeared as states in the Markov law for W or D. In such instances, Lucas and Prescott's social planning problem fails to reproduce the rational expectations equilibrium essentially because the fictitious planner takes into account the externality that the feedback from K to W or D constitutes.

[18] This was emphasized in a more general context by Lucas and Prescott (1971).

subject to the laws of motion (8), (4), (5), and (2) for w_t, u_t, D_{1t}, and ϵ_t[19] and subject to K_0 given.

The term in brackets is the area under the demand curve, since

$$\int_0^{Y_t} (A_0 - A_1 x + A_2 D_{1t} + u_t) dx = A_0 Y_t - \frac{1}{2} A_1 Y_t^2 + Y_t A_2 D_{1t} + Y_t u_t.$$

Thus (17) is the discounted area under the demand curve minus the total costs of production. Dividing each side of (16) by βd, the Euler equation can be written

$$\begin{aligned} K_{t+1} - \left(1 + \frac{1}{\beta} + \frac{A_1 f^2 n}{d}\right) K_t + \frac{1}{\beta} K_{t-1} \\ = \frac{n}{d} w_t - \frac{nfA_0}{d} + \frac{A_1 fn}{d} \epsilon_t \\ - d^{-1} fnA_2 D_{1t} - \frac{fn}{d} u_t. \end{aligned} \tag{18}$$

It can easily be proved that there exists a λ such that

$$[1 - (1 + \beta^{-1} + A_1 f^2 n d^{-1}) L + \beta^{-1} L^2] = [1 - (\lambda\beta)^{-1} L](1 - \lambda L),$$

where $|\lambda| < 1/\sqrt{\beta}$.[20] Using $-(\lambda\beta)^{-1} L(1 - \lambda\beta L^{-1}) = [1 - (\lambda\beta)^{-1} L]$, we have that the Euler equation (18) can be written as

$$\begin{aligned} -(\lambda\beta)^{-1} L (1 - \lambda\beta L^{-1})(1 - \lambda L) K_{t+1} = \frac{-nfA_0}{d} + \frac{n}{d} w_t \\ + \frac{A_1 fn}{d} \epsilon_t - \frac{fn}{d} A_2 D_{1t} - \frac{fn}{d} u_t. \end{aligned}$$

A solution of the Euler equation that also satisfies the transversality condition for the social planning problem is

$$\begin{aligned} (1 - \lambda L) K_{t+1} = \frac{+\lambda\beta nfA_0}{d} (1 - \lambda\beta)^{-1} - \frac{n\lambda\beta}{d} \frac{L^{-1}}{1 - \lambda\beta L^{-1}} w_t \\ - \frac{A_1 fn\lambda\beta}{d} \frac{L^{-1}}{1 - \lambda\beta L^{-1}} \epsilon_t + \frac{fn\lambda\beta d^{-1} L^{-1}}{1 - \lambda\beta L^{-1}} A_2 D_{1t} \\ + \frac{fn\lambda\beta d^{-1} L^{-1}}{1 - \lambda\beta L^{-1}} u_t. \end{aligned} \tag{19}$$

Recall, for example, that $(1 - \lambda\beta L^{-1})^{-1} w_t = \sum_{j=0}^{\infty} (\lambda\beta)^j w_{t+j}$. Then it can be recognized that equation (19) is the perfect foresight solution of

[19] It can also be proved that the transversality condition for (17) imposes the same condition on the solution as does the transversality condition of the representative firm.

[20] This follows directly from the observation that if z_0 is a zero of $[1 - (1 + \beta^{-1} + A_1 f^2 n d^{-1}) z + \beta^{-1} z^2]$, then so is βz_0^{-1}.

the planning problem that the rational expectations competitive equilibrium implicitly solves. Thus, equation (19) expresses the aggregate capital stock K_{t+1} as a linear function of K_t and all future values of w_t, ϵ_t, D_{1t}, and u_t.

By using the methods of Hansen and Sargent (1980*b*, esp. appendix A), equation (19) can be converted to the "realizable" law for K that satisfies the Euler equations and transversality conditions, and which expresses K_{t+1} as a function only of information known at time t.[21] This involves replacing the terms w_{t+i}, ϵ_{t+i}, D_{1t+i}, and u_{t+i} in (19) by the corresponding mathematical expectations conditioned on Ω_t. The resulting equilibrium law of motion for K can be shown to be

$$\begin{aligned} K_{t+1} = H_0 + H_w(L)W_t + H_D(L)D_t + H_\epsilon(L)\epsilon_t \\ + H_u(L)u_t + H_1K_t, \end{aligned} \tag{20}$$

where

$$\begin{aligned} H_0 &= \frac{+\lambda\beta nfA_0}{d(1-\lambda\beta)}, \\ H_1 &= \lambda, \\ H_w(L) &= \frac{-n\lambda\beta}{d}\phi_w\left\{\frac{L^{-1}[I-\delta_w(\lambda\beta)^{-1}\delta_w(L)]}{1-\lambda\beta L^{-1}}\right\}, \\ H_\epsilon(L) &= \frac{-A_1fn\lambda\beta}{d}\left\{\frac{L^{-1}[1-\delta_\epsilon(\lambda\beta)^{-1}\delta_\epsilon(L)]}{1-\lambda\beta L^{-1}}\right\}, \\ H_D(L) &= +fn\lambda\beta d^{-1}A_2\phi_D\left\{\frac{L^{-1}[I-\delta_D(\lambda\beta)^{-1}\delta_D(L)]}{1-\lambda\beta L^{-1}}\right\}, \\ H_u(L) &= +d^{-1}fn\lambda\beta\left\{\frac{L^{-1}[1-\delta_u(\lambda\beta)^{-1}\delta_u(L)]}{1-\lambda\beta L^{-1}}\right\}. \end{aligned} \tag{21}$$

Here ϕ_w is a $1 \times q$ vector with 1 in the first position, followed by $(q-1)$ zeroes, and ϕ_D is a $p_1 \times p$ matrix with a $(p_1 \times p_1)$ identity matrix as the first p_1 columns and zeroes elsewhere. Notice that $w_t \equiv \phi_w W_t$ and $D_{1t} \equiv \phi_D D_t$. It is convenient at this point to recall the laws of motion assumed for w_t, u_t, ϵ_t, and D_{1t}, namely,

$$\delta_w(L)W_t = V_t^w, \tag{8}$$

$$\delta_u(L)u_t = V_t^u, \tag{4}$$

[21] Note that eq. (19) satisfies the first-order necessary conditions for the optimization problem but gives the planner too much information (it is "anticipative" or "nonrealizable"). The correct solution to the problem taking the information set available to the planner into account is the solution of the first-order necessary conditions that expresses K_{t+1} as a function only of information that the planner possesses at time t. Such a solution is said to be "realizable" or "nonanticipative."

$$\delta_D(L)D_t = V_t^D, \tag{5}$$

$$\delta_\epsilon(L)\epsilon_t = V_t^\epsilon. \tag{2}$$

Equation (20) expresses the equilibrium law for the industry-wide capital stock K_{t+1} as a linear function of K_t and current and past values of W, D, ϵ, and u. Current and past values of W appear in (20) because they help predict future values of the rental rate w_t, while current and past values of D, ϵ, and u appear because they are used by agents to predict the future course of the market price P. The numbers of lagged values of W, D, ϵ, and u in (20) are $r_w - 1$, $r_D - 1$, $r_\epsilon - 1$, and $r_u - 1$, as expressions (21) can be used to show.[22] Thus, the numbers of lagged values of these "information variables" W, D, ϵ, and u in (20) are entirely inherited from the specifications of the actual laws of motion for W, D, ϵ, and u in (8), (5), (2), and (4).

Notice that the appearance of w_t and D_{1t} in the objective function of the representative firm (9) (or equivalently in the objective function of the fictitious social planner [17]) gives rise to the appearance in (20) of the entire blocks of variables W_t and D_t that help predict w and D_{1t}, respectively. Thus any variables that help predict w and D_{1t}, and which agents have information on, belong in the equilibrium law of motion for industry-wide capital. The property that the remaining variables in W (or D) help predict future values of w (or D_1) is said to be the property that the remaining variables in W (or D) *Granger cause* w (or D_1). The notion of Granger causality thus turns out to be coincident with the criterion for whether random variables that do not themselves appear in the agent's criterion function nevertheless end up in the equilibrium law of motion or decision rule, essentially because they appear in the agents' constraints as information variables that help predict variables that do appear in the criterion function. It is mainly for this reason that the concept of Granger causality has played an important role in work with rational expectations models.[23]

[22] By expanding the polynomial in L, it is possible to show that

$$\frac{L^{-1}[I - \delta(\lambda\beta)^{-1}\delta(L)]}{1 - \lambda\beta L^{-1}} = \delta(\lambda\beta)^{-1}\left\{\sum_{j=0}^{r-1}\left[\sum_{k=j+1}^{r}(\lambda\beta)^{k-j-1}\delta_k\right]L^j\right\}, \tag{22}$$

where $\delta(L) = I - \sum_{j=1}^{r}\delta_j L^j$. Notice that the polynomial on the left side of (22) is one sided in nonnegative powers of L, despite the appearance of L^{-1}, and that it is a polynomial of order $(r - 1)$, as asserted in the text. The formula (22) can be derived by mimicking the procedures used in Hansen and Sargent (1980*a*). The same mathematical techniques used by Hansen and Sargent (1980*a*) to derive expressions like (21) or (22) were independently utilized by Futia (1979) to compute linear rational expectations equilibria. Also, without knowing of Hansen and Sargent's work, John Kennan independently derived formulas similar to (22) in a personal letter to me.

[23] From this point of view, it is irrelevant whether Granger causality is consistent with one's notion of what "true" causality is. Sims (1972) has described the relationship of the concept of Granger causality to that of strict econometric exogeneity. That a random process y fail to Granger cause x is a necessary condition for x to be strictly econometri-

Equations (20) and (21) reveal explicitly how the parameters of the equilibrium law of motion for industry-wide capital are themselves nonlinear functions of the underlying parameters $\{A_0, A_1, A_2, f, d, \beta, n, \delta_w(L), \delta_u(L), \delta_\epsilon(L), \delta_D(L)\}$. The nonlinearity has two sources. First, there is the fact that λ is a nonlinear function of β and $(A_1 f^2 nd^{-1})$ via the factorization defining λ, $[1 - (\lambda\beta)^{-1}L](1 - \lambda L) = [1 - (1 + \beta^{-1} + A_1 f^2 nd^{-1})L + \beta^{-1}L^2]$. Second, given λ, the formulas for $H_w(L), H_u(L), H_\epsilon(L)$, and $H_D(L)$ in (21) are nonlinear in the parameters of $\delta_w(L)$, $\delta_u(L)$, $\delta_\epsilon(L)$, and $\delta_D(L)$. Nonlinear cross-equation restrictions of the kind illustrated by (20) and (21) are the hallmark of rational expectations models. Such cross-equation restrictions are largely absent from "pre–rational expectations" dynamic econometric models.[24] The presence of these restrictions impinges on a variety of fundamental econometric and conceptual issues, including identification, the analysis of interventions, models of "error terms," and the role of "prior information." I now turn to discussing each of these issues, using (20) and (21) as an instrument.

Analysis of Interventions

At this point, it is useful to remind ourselves of the principal reason that an economist might want to construct a dynamic econometric model of an industry along the lines of our example. It is to be able to make quantitative predictions about the effects on the industry that various hypothetical interventions or "changes in the environment" will have. In the present context, a hypothetical "intervention" or "change in the environment" means a change in one of the polynomials $\delta_w(L)$, $\delta_u(L)$, $\delta_D(L)$, or $\delta_\epsilon(L)$ that describe, respectively, the stochastic processes for W, u, D, and ϵ that impinge on the market.[25] Several interesting examples of such interventions can be given, including the following:

a) Suppose that there is a specific tax imposed on sales of the

---

cally exogenous with respect to y. For this reason, the concept of Granger causality is also useful in designing specification tests. For a discussion of the relationship between Granger causality and econometric exogeneity in the context of linear rational expectations models, see Hansen and Sargent (1980*a*).

[24] Thus Fisher wrote: "In practice, except for such covariance restrictions [across disturbances in distinct structural equations], restrictions which relate the parameters of one equation to those of one or more others are extremely rare. There is no reason in principle why such cases cannot occur, however, and it may be worthwhile devoting a very short discussion to them" (1966, p. 176).

[25] By now, this is a routine and uncontroversial definition of an intervention. Applications of the techniques of optimal control theory to the calculation of macroeconometric and microeconometric policy response functions employ precisely this concept of intervention (see, e.g., Chow 1973; Kareken, Muench, and Wallace 1973; Arzac and Wilkinson 1979; and Taylor 1979).

product. Such a specific tax can be modeled as a component of $(A_2 D_{1t})$. Since the behavior of the tax through time will be described by an element of the vector Markov law $\delta_D(L)D_t = V_t^D$, changes in the rule for setting the specific tax amount to changes in one of the rows of $\delta_D(L)$.

b) Suppose that there is a specific tax on the use of the factor of production. This tax can be modeled as an addition to the rental w_t. A change in the rule for setting this tax can be modeled as a change in a row of $\delta_w(L)$.

c) Suppose there is a change in the structure of the process governing the "pretax" part of the rental. Again this can be modeled as a change in one row of $\delta_w(L)$. With a little imagination, the effects of a change in the organization of the industry[26] supplying the factor might be modeled in this way.

The model leading to (20) and (21) provides a way of predicting quantitatively the effects of such changes, once agents have caught on to them. The effect of interventions in the sense described here is to change the function (20) describing the evolution of industry capital in a way predicted by the formulas given in (21). Since interventions of this class change the law of motion (20), it is necessary to have analytic methods which use the cross-equation restrictions (21) to predict how the H's of the K-law of motion (20) will change if there is a hypothetical intervention operating on one or more of the δ's.

In order to evaluate policy interventions in this way, it is essential that the H's of (20) should not be viewed as being among the free parameters of the model. Instead, the model's free parameters are to be regarded as the deeper parameters $\{A_0, A_1, A_2, f, d, \delta_w, \delta_u, \delta_\epsilon, \delta_D\}$. The researcher needs to know these parameters in order to be able to use the formulas (21) to predict the consequences of hypothetical changes in the functions δ.[27]

From the dynamic economic theory leading to (20) and (21), it is

[26] E.g., if it becomes a cartel when before it had been competitive or noncooperative in some way.

[27] A technical qualification needs to be added at this point. In order to have a model capable of predicting effects of interventions acting on the δ's, one can sometimes get by without having uniquely identified the parameters $\{A, f, d\}$. What the researcher must identify are the parameters of the characteristic polynomial of the Euler equation (16), namely the parameters ϕ_0, ϕ_1 in $\{\beta d - [d(1+\beta) + A_1 f^2 n\beta]L + dL^2\} = (\beta\phi_1 + \phi_0 L + \phi_1 L^2)$, where $\phi_1 = d$ and $-\phi_0 = [d(1+\beta) + A_1 f^2 n\beta]$. The theory assumes that the ϕ_j's will be invariant with respect to interventions on the δ's. If the researcher can uniquely identify the ϕ_j's, he can proceed with econometric policy evaluation, even if he cannot uniquely identify all of (A_1, f, d). In some setups, the parameters of the characteristic polynomials of the Euler equations are identified even though only an equivalence class of the counterparts of (A_1, f, d) is identified. This is enough for econometric policy evaluation to proceed. This problem is discussed by Hansen and Sargent (1980*b*). It is technically related to the "inverse optimal control" problem (see Mosca and Zappa 1979).

evident that a given numerical version of (20), estimated from historical data, cannot be used to evaluate the consequences of arbitrary input sequences for $\{W_t\}$, $\{D_t\}$, $\{\epsilon_t\}$, and $\{u_t\}$. That is, a fixed law of motion of the form (20) with given numerical values for the H's cannot be used to investigate the consequence of arbitrarily specified numerical sequences for the W, D, ϵ, and u's. In effect, a particular version of (20) can be expected to hold up only for W, D, ϵ, and u sequences drawn from a restricted domain: namely, sequences obeying the probability laws (8), (4), (5), and (2).[28]

However, until Lucas wrote in 1976, evaluating the effects of interventions in this inappropriate way was the accepted procedure in both the macroeconometric and the microeconometric literatures. Regrettably, to this day it remains the procedure used in the overwhelming majority of analyses of policy interventions. It should be emphasized once again that from the viewpoint of the dynamic decision theory described above, the question of how agents will respond to "arbitrary sequences" of "forcing variables" W, u, ϵ, and D is not well posed. In effect, unless the researcher specifies precisely the perceived laws of motion for the "forcing variables," he has not specified the constraints subject to which decision makers are thought to be acting.[29]

Thus, in order to be able to evaluate interventions operating on the δ's, it is necessary to formulate and estimate the model in terms of the parameters of preferences (A_0, A_1, A_2), technology (f and d), and the constraints (the δ's). The argument in favor of formulating and estimating the dynamic model at the level of the deep parameters $\{A_0, A_1, A_2, d, f, \delta_w, \delta_u, \delta_\epsilon, \delta_D\}$ is in much the same spirit as the usual justification for estimating "structural" parameters rather than reduced-form parameters. As Marschak (1953) argued, the researcher wants to estimate those objects which will permit him to analyze an interesting class of changes in the environment. Dynamic economic theory has forced us to reexamine whether objects long thought to be "structural," including the parameters of decision rules such as consumption, investment, and portfolio balance schedules, are correctly taken to be invariant with respect to changes in the environment. Once agents' behavior is modeled in terms of genuinely dynamic optimization problems, it becomes apparent that the parameters of observed decision rules should not be viewed as structural (see

[28] This message is at least implicit in the work by Lucas and Prescott (1971). Gordon and Hynes (1970) made the argument in an informal way. Lucas (1976) forcefully brought the message to the attention of macroeconomists.

[29] However, some economists continue to argue that existing macroeconometric models can be used to predict the effects of such arbitrary sequences (see Friedman 1978).

Muth 1961; Lucas and Prescott 1971; Merton 1971; and Lucas 1972*a*).

The Neglect of Learning

At this point it is worthwhile to discuss a modification of the preceding kind of setup which several economists have apparently had in mind.[30] For this purpose it is sufficient to consider the problem of maximizing the social welfare criterion subject to the given laws of motion (8), (4), (5), and (2) for W_t, u_t, D_t, and ϵ_t. By relabeling and reinterpreting the variables, we can think of this as a choice problem faced by a single private agent. In posing this problem, it was assumed that the agent solving the problem knows the true values of the parameters of the objective function (17) and the true values of the polynomials in the lag operator $\delta_\epsilon(L)$, $\delta_u(L)$, $\delta_D(L)$, and $\delta_w(L)$. The observation has been made that this setup fails to incorporate a model of how the agent optimally learns about the δ's from observations on past realizations of the forcing variables ϵ, u, D, and W. Presumably, if the agent has only finite histories of observations on ϵ, u, D, and W at his disposal, then at each point in time he is uncertain about the parameters of the polynomials δ. Why not modify the preceding setup to include uncertainty about the δ's and a model of optimal learning about the δ's? There seem to be three reasons why such extensions have not as yet successfully been incorporated into rational expectations models.

The first is as follows. A general model of optimal learning about the δ's is readily available in the "Kalman filter," which can be used to model how a rational agent would use observations on $(\epsilon_t, u_t, D_t, W_t)$ to revise his prior beliefs about the δ's.[31] However, with the δ's uncertain, it is no longer possible to give closed-form formulas for the optimal decision rule in terms of what are now the posterior probability distributions over the δ's. The reason that no one has yet obtained or is likely ever to obtain such closed formulas is as follows. In deriving the closed form of the restrictions (21) for the case in which the δ's are

[30] See Friedman 1979 and Modigliani 1977.

[31] See Anderson and Moore 1979. The Kalman filter provides a model of Bayesian learning about the δ's where the initial prior and the posteriors are multivariate normal. However, as Hansen points out to me, normal posteriors for the δ's are inadmissible for dynamic models of the class described here. This is because the dynamic optimization problems we consider may be ill posed for points in the parameter space of δ's for which the zeroes of det $\delta(z)$ are less than $\sqrt{\beta}$ in modulus. Only priors and posteriors that assign zero probability to this region of the parameter space are in general admissible for our problems. This rules out multivariate normal distributions. Taking account of this admissibility constraint severely complicates the task of building a model of optimal learning about the δ's.

assumed known with certainty, the Wiener-Kolmogorov prediction formula,

$$E_t W_{t+i} = \left[\frac{\delta_w(L)^{-1}}{L^i}\right]_+ \delta_w(L) W_t,$$

was used extensively. Here $[\sum_{j=-\infty}^{\infty} \alpha_j L^j]_+ = \sum_{j=0}^{\infty} \alpha_j L^j$, so that $[\quad]_+$ means "ignore negative powers of L." The Wiener-Kolmogorov formula is equivalent with the "chain rule" of forecasting (see Shiller [1972] or Sargent [1979] for expositions). These equivalent forecasting rules are known to be correct for the case in which the δ's are known with certainty. However, as Chow (1973) has pointed out, where there is a nontrivial posterior density over the δ's, there is in general no known closed-form formula such as the above one for the i-step-ahead forecast. For example, it is not true that where δ is uncertain, the correct expression for $E_t W_{t+i}$ is given by replacing the δ_{wj}'s with their posterior means in the above formula. The fact that there is no closed-form prediction formula for sufficiently general cases implies that it is impossible to derive closed-form versions of decision rules (and hence equilibria) that correspond to (21). As we shall see, for the kind of empirical work we are advocating, it is important to have a closed form for the mapping from the parameters of the objective functions (17) and the dynamic constraints to the decision rule (20). From this viewpoint, the suggestion that one ought to build a learning mechanism into rational expectations models is not useful in suggesting practical econometric alternatives to the procedures recommended here.[32]

Another drawback with incorporating learning is that, even if one could derive the decision rules in the face of uncertain δ's, the issue would arise of how to determine the prior used to initiate the learning model for the δ's. Would it be imposed a priori or estimated? If the initial prior were to be estimated, this would substantially complicate the estimation problem and add to the number of parameters.

Finally, in many settings the Bayesian learning model implies that the posterior distributions collapse about the true δ's as time passes without limit. In such settings, even if the researcher erroneously assumes that the δ's are known with certainty when in reality agents are learning about them in an optimal way, the researcher continues to obtain consistent estimators of the underlying parameters $\{A_0, A_1,$

[32] Further, notice that if the decision rules could be calculated in closed form under uncertainty about the δ's, the resulting time-series models would have time-varying coefficients and so be nonstationary. Even if calculating the decision rules were a tractable task under uncertainty about the δ's, the loss of stationarity that it would imply might well be a price that the applied economist would not be prepared to pay even in exchange for the "greater realism" of the learning assumption.

$A_2, f, d, \delta_w, \delta_D, \delta_\epsilon, \delta_u\}$ using the methods described here and in Hansen (1979) and Hansen and Sargent (1980*a*). It does seem likely that by erroneously ignoring the phenomenon of learning about the δ's, the researcher is incorrectly calculating the asymptotic covariance matrix of his estimators. However, at present nothing is known about the nature of this error. Further, since we simply do not know how to compute optimum decision rules under the assumption that agents know the δ's with uncertainty, no consistent estimators of the underlying parameters have been proposed that incorporate agents' learning about the δ's in the optimal way, to say nothing of expressions for the associated asymptotic covariance matrices.

From the preceding considerations, I draw the conclusion that incorporating optimal Bayesian learning about the δ's on the part of agents is not a research avenue that soon promises appreciable dividends for the economist interested in applying dynamic competitive models of the sort described here.

A Model of the "Error Term"

We now derive a "dynamic supply curve" for the industry by using the industry-wide production function $Y_t = fK_t + \epsilon_t$ to eliminate K from (20) in favor of Y. Multiplying both sides of (20) by f and then adding ϵ_{t+1} to both sides gives

$$\begin{aligned} Y_{t+1} = H_0 f &+ fH_w(L)W_t + fH_D(L)D_t \\ &+ fH_\epsilon(L)\epsilon_t + fH_u(L)u_t \\ &+ H_1Y_t + \epsilon_{t+1} - H_1\epsilon_t. \end{aligned}$$

Eliminating u_t by using $u_t = P_t - A_0 + A_1Y_t - A_2D_{1t}$ gives

$$\begin{aligned} Y_{t+1} = [H_0 f - fH_u(1)A_0] &+ fH_u(L)P_t + fH_w(L)W_t \\ &+ [fH_D(L) - fH_u(L)A_2\phi_D]D_t \\ &+ [H_1 + fH_u(L)A_1]Y_t \\ &+ [1 + fH_\epsilon(L)L - H_1L]\epsilon_{t+1}. \end{aligned} \tag{23}$$

This can be written as

$$\begin{aligned} Y_{t+1} = S_0 &+ S_p(L)P_t + S_w(L)W_t \\ &+ S_D(L)D_t + S_Y(L)Y_t + S_\epsilon(L)\epsilon_{t+1}, \end{aligned} \tag{24}$$

where

$$\begin{aligned} S_0 &= H_0 f - fH_u(1)A_0, \\ S_p(L) &= fH_u(L), \\ S_w(L) &= fH_w(L), \\ S_D(L) &= [fH_D(L) - fH_u(L)A_2\phi_D], \\ S_Y(L) &= [H_1 + fH_u(L)A_1], \\ S_\epsilon(L) &= [1 + fH_\epsilon(L)L - H_1L]. \end{aligned} \tag{25}$$

Recall that the demand curve is

$$P_t = A_0 - A_1Y_t + A_2D_{1t} + u_t. \tag{3}$$

Using $\delta_\epsilon(L)\epsilon_t = V_t^\epsilon$ and $\delta_u(L)u_t = V_t^u$, we can write the supply and demand curves as

$$\begin{aligned} Y_{t+1} = S_0 &+ S_p(L)P_t + S_w(L)W_t + S_D(L)D_t \\ &+ S_Y(L)Y_t + S_\epsilon(L)\delta_\epsilon(L)^{-1}V_{t+1}^\epsilon, \end{aligned} \tag{26}$$

$$P_t = A_0 - A_1Y_t + A_2D_{1t} + \delta_u(L)^{-1}V_t^u. \tag{27}$$

To discuss identification and estimation of the model, we need a theory about what is unknown to the econometrician. In constructing the model, we have taken the view that all of the variables on the right-hand side of the supply and demand curves (24) and (3) (or equivalently [26] and [27]) are known to the representative firm. Thus, from the viewpoint of private agents, (26) and (27) describe exact linear functions of the right side variables in which there are no "random errors."[33] The only tractable way that has so far been discovered of introducing random errors into (26) and (27) has been to assume that the econometrician has less information than do the private agents. The smaller information set of the econometrician leads to what from his point of view are random terms in relationships to be derived from (24) and (3) or (26) and (27). The idea is to restrict the econometrician's information set relative to that of private agents in a way both that is plausible and that leads to a tractable statistical model of the error term. I shall describe two models of the error term that can be constructed in this way.

One model results from assuming that the econometrician has time series on $\{P_t, W_t, D_t, Y_t\}$ but never observes the random processes ϵ_t and u_t. On this interpretation, ϵ_t and u_t become random terms in (24)

[33] This is a consequence of the fact that the representative firm views itself as playing a dynamic "game against nature," and so finds it optimal to use a nonrandom strategy, that is, a strategy that can be expressed as an exact function of its information variables and other state variables.

and (3) from the econometrician's viewpoint.[34] In constructing the model, we have already imposed that V^{ϵ}_{t+1} is orthogonal to all of the variables on the right side of (26) and will assume that V^{u}_{t} is orthogonal to Y_t. We can also impose that V^{u}_{t} is orthogonal to D_{1t}, if we wish,[35] although we might get by with a weaker assumption.

The second model of the error term results from assuming that the econometrician sees less of D_t and W_t than do private agents. It is convenient to postpone a detailed discussion of this second model of the error and, instead, first to discuss identification and estimation under the first model of the error term.

Identification and Estimation

With this model of the error terms, we can proceed to discuss identification and estimation. First, notice that every variable that appears on the right side of the demand schedule (27) also appears on the right side of the supply schedule (26). The dynamic economic theory leading to (26) makes the reason for this clear, since any variables that help predict future prices P will appear in the supply schedule of the representative firm. This immediately implies that any variables that help predict the demand shifters D_{1t} will appear in the supply schedule.[36] The fact that no variables on the right side of the demand curve (27) are excluded from the supply schedule (26) means that, if the supply schedule is to be identified, the source of identification must be found in restrictions of a kind different from the usual exclusion restrictions treated extensively in econometrics textbooks.[37] According to the standard "order condition" for identification, equa-

[34] This is a version of the model of the error term analyzed by Hansen and Sargent (1980*a*) and Sargent (1978).

[35] We have assumed that V^{ϵ}_{t} and V^{u}_{t} are the "innovations" or one-step-ahead errors in predicting ϵ_t and u_t on the basis of observations on all variables in agents' information set at time $t - 1$ (see n. 6). This implies that V^{ϵ}_{t+1} is orthogonal to all variables on the right side of (26). If we assume that V^{u}_{t} is orthogonal to V^{ϵ}_{t}, it also implies that V^{u}_{t} is orthogonal to Y_t. Imposing that V^{u}_{t} is orthogonal to D_{1t} amounts to assuming that D_{1t} is strictly exogenous in (27), which is stronger than the Granger causality assumptions already imposed on D_{1t}, namely, that except for lagged D's, no other variables in the model Granger cause D_{1t}.

[36] There is a singular class of exceptions to this statement. In the special case that

$$ED_{1t+j} \mid \{D_{t-s}\}_{s=0}^{\infty} = ED_{1t+j} \mid \{D_{2t-s}\}_{s=0}^{\infty} \tag{28}$$

for all $j \geq 1$, D_{1t}'s will appear in the demand schedule but not in the supply schedule. The condition (28) is usually thought to be exceedingly unlikely for any economic time series $\{D_{1t}\}$.

[37] These remarks about identification should be compared with Milton Friedman's discussion (1953) of the conditions needed for "supply" and "demand" to provide a useful categorization of the factors impinging on price and output. Friedman argued that the categorization was useful to the extent that it effectively sorted forces acting on price and output into mutually exclusive categories.

tion (26) is hopelessly underidentified.[38] Thus if the parameters of the model are to be identified, sources of prior information not of the exclusion variety must be available. The main source of these restrictions in the present model is the extensive body of cross-equation restrictions embodied in equations (21) and (25). Equations (21) and (25) give the parameters of the supply schedule (26) as nonlinear functions of the parameters $\{A_0, A_1, A_2, f, d, \beta, n, \delta_w(L), \delta_D(L), \delta_u(L),$ and $\delta_\epsilon(L)\}$. In general, provided that the parameters r_D and r_w, which determine the order of the autoregressive processes for D and W, and the parameters p and q, the number of elements in the vectors D and W, respectively, are large enough, these cross-equation restrictions identify or overidentify the parameters of the model. The strength of overidentification generally increases with increases in the orders r_D and r_w and the dimensions p and q.[39]

At this point it is useful to collect together the equations comprising the model as

$$\begin{aligned} Y_{t+1} = S_0 + S_p(L)P_t + S_w(L)W_t + S_D(L)D_t \\ + S_Y(L)Y_t + S_\epsilon(L)\delta_\epsilon(L)^{-1}V_{t+1}^{\epsilon}, \end{aligned} \tag{26}$$

$$P_t = A_0 - A_1 Y_t + A_2 D_{1t} + \delta_u(L)^{-1}V_t^u, \tag{27}$$

$$\delta_w(L)W_t = V_t^w, \tag{8}$$

$$\delta_D(L)D_t = V_t^D, \tag{5}$$

where

$$\begin{aligned} S_0 &= +f^2\lambda n\beta A_0 d^{-1} - fH_u(1)A_0, \\ S_p(L) &= +f^2 n\lambda\beta d^{-1}\left\{\frac{L^{-1}[I - \delta_u(\lambda\beta)^{-1}\delta_u(L)]}{1-\lambda\beta L^{-1}}\right\}, \\ S_w(L) &= \frac{-n\lambda\beta f}{d}\phi_w\left\{\frac{L^{-1}[I - \delta_w(\lambda\beta)^{-1}\delta_w(L)]}{1-\lambda\beta L^{-1}}\right\}, \\ S_D(L) &= +f^2 n\lambda\beta d^{-1}A_2\phi_D\left\{\frac{L^{-1}[I - \delta_D(\lambda\beta)^{-1}\delta_D(L)]}{1-\lambda\beta L^{-1}}\right\} \\ &\quad - f^2 d^{-1} n\lambda\beta\left\{\frac{L^{-1}[I - \delta_u(\lambda\beta)^{-1}\delta_u(L)]}{1-\lambda\beta L^{-1}}\right\}A_2\phi_D, \end{aligned} \tag{29}$$

[38] The fact that the demand curve excludes some variables that appear in the supply schedule is due to the static specification for the demand curve. This feature of the model would not survive a variety of alterations that might plausibly be used to introduce dynamics into the demand curve. E.g., if the demand schedule involved expected future prices as arguments, all variables that help to predict future prices would appear in the econometrically operational expression for current P_t that would correspond to (27).

[39] This characteristic of identification in rational expectations models has been noted in various contexts by several authors, including Lucas (1975) and Sims (1980).

$$S_Y(L) = +f^2 d^{-1} n\lambda\beta A_1\left\{\frac{L^{-1}[I - \delta_u(\lambda\beta)^{-1}\delta_u(L)]}{1 - \lambda\beta L^{-1}}\right\} + \lambda,$$

$$S_\epsilon(L) = 1 - \frac{A_1 f^2 n\lambda\beta}{d}\left[\frac{I - \delta_\epsilon(\lambda\beta)^{-1}\delta_\epsilon(L)}{1 - \lambda\beta L^{-1}}\right] - \lambda L.$$

$$[1 - (\lambda\beta)^{-1}L](1 - \lambda L) = [1 - (1 + \beta^{-1} + A_1 f^2 n d^{-1})L + \beta^{-1}L^2].$$

Equations (26), (27), (8), and (5) form a statistical model for the joint process (P_t, Y_t, W_t, D_t). The model is linear in the variables but is characterized by the extensive set of cross-equation restrictions described by (29). With the model of the error terms currently under discussion, the statistical model of the (P_t, Y_t, W_t, D_t) process has been spelled out sufficiently completely that we could write down the likelihood function for a sample (P_t, Y_t, W_t, D_t), $t = 1, \ldots, T$, assuming a normal probability density for $(V_t^w, V_t^u, V_t^D, V_t^\epsilon)$.[40] Maximum likelihood estimates of the free parameters of the model $\{A_0, A_1, A_2, f, d, \delta_D(L), \delta_w(L), \delta_u(L), \delta_\epsilon(L)\}$ could then be obtained. Computational details of such procedures are described by Sargent (1977, 1978) and Hansen and Sargent (1980*a*). From the point of view of computing the estimates, it is a great practical advantage that (29) gives a set of closed-form formulas for the cross-equation restrictions imposed by the dynamic economic theory.

Application of Bayesian Methods

The fact that for the present model of the error terms it is possible to write down a normal likelihood function means that in principle Bayesian methods are applicable. Letting θ be the list of parameters of the model and Z be the data, we have

$$f_{\text{post}}\{\theta \mid Z\} = \frac{l\{Z \mid \theta\} f_{\text{prior}}\{\theta\}}{f(Z)},$$

or

$$f_{\text{post}}\{\theta \mid Z\} = l(Z \mid \theta) f_{\text{prior}}(\theta) / \int l(Z \mid \theta) f_{\text{prior}}(\theta) d\theta, \quad (30)$$

where $f_{\text{post}}\{\theta \mid Z\}$ denotes the posterior probability density, $f(Z)$ the probability density of Z, $f_{\text{prior}}\{\theta\}$ the prior density on θ, and $l\{Z \mid \theta\}$ the likelihood function. Measures of the location and dispersion of the posterior distribution of θ can be calculated, for example, by integrating $\theta^k \cdot f_{\text{post}}\{\theta \mid Z\}$ over θ for appropriate values of k. In the Bayesian view, the role of data analysis is to trace out in as revealing a way as possible the mapping defined by (30) from the prior to the

[40] See Hansen and Sargent (1980*a*) for a discussion of the details.

posterior distribution. For such an analysis to be practical, it substantially eases matters if the mapping (30) can be characterized analytically so that, for example, posterior moments such as $\int \theta f_{\text{post}}(\theta \mid Z) d\theta$ can be calculated without the need to resort to numerical integration. Zellner (1971) and Leamer (1978) describe forms of prior densities $f_{\text{prior}}(\theta)$ that have the property that the mapping (30) is one that can be written as an analytic closed form when $l(Z \mid \theta)$ is the normal likelihood function.

In the context of dynamic economic models of the class represented by (26), (27), (8), and (5), the question of whether the mapping (30) can be characterized analytically hinges on which parameters one regards as being in the list θ about which the researcher has formulated prior information. One possibility is that θ consists of the S's of (26), the A's of (27), and δ_w, δ_u, and δ_D of (27), (8), and (5). With this interpretation of θ, then since (26), (27), (8), and (5) are linear in the S's, A's, δ_w, and δ_D, it is possible to get analytic characterizations of the mapping from $f_{\text{prior}}(\theta)$ to $f_{\text{post}}(\theta \mid Z)$. For example, Leamer (1972) and Shiller (1972) have shown how priors of various forms on the S's in (26) can tractably be mapped into posteriors, in contexts where (26) is appropriately viewed as a regression equation. In effect, Leamer (1972) and Shiller (1973) provided formal Bayesian methods for imposing restrictions on lag distributions of a general kind, examples of which had long been imposed by applied econometricians. These restrictions usually corresponded to restrictions directly on our S_j's. Predating the work of Shiller and Leamer were the restrictions on distributed lags proposed by Koyck (1954), Cagan (1956), Milton Friedman (1957), Almon (1965), and Jorgenson (1966). There was also the frequently used identifying restriction that various distributed lag weights sum to unity.[41] All of these approaches view the S's themselves as among the free parameters of the model about which the researcher can reasonably be imagined to have formed views summarized by a prior distribution.

Unfortunately, the tractability of the Leamer-Shiller approach is purchased at the cost of ignoring the essential aspects of the dynamic economic theory leading to (26). According to that theory, the S's are not free parameters but are complicated functions of the parameters $\{A_0, A_1, A_2, f, d, \beta, n, \delta_w(L), \delta_D(L), \delta_u(L), \delta_\epsilon(L)\}$. It is this list of parameters about which it seems most appropriate to expect an economist to have prior beliefs. The parameters $\{A_0, A_1, A_2, f, d\}$ are the parameters describing preferences and the technology, about which the economic theorist may have some prior beliefs. The

[41] This restriction was criticized by Lucas (1972*b*) and Sargent (1971) for essentially the same reasons given here.

economists' prior beliefs about the parameters $\{\delta_w, \delta_D, \delta_u, \delta_\epsilon\}$ are presumably on a different theoretical footing than his beliefs about $\{A_0, A_1, A_2, f, d\}$, since the former list simply characterizes the serial correlation properties of the "shift variables" about which economic theory itself suggests little, although casual general observations may suggest a presumption in favor of high serial correlation, at least in some types of variables. In any event, it is the deep parameters $\{A_0, A_1, A_2, f, d, \beta, n, \delta_w, \delta_D, \delta_u, \delta_\epsilon\}$ that must be estimated, if one is to build a model that potentially overcomes Lucas's critique of econometric policy evaluation procedures.[42]

When this list of deep parameters contains the objects of interest, Bayesian analysis using (30) becomes much less tractable. This is because the likelihood function $l(Z \mid \theta)$ becomes a very complicated function of the free parameters in θ, by virtue of the complicated nature of the cross-equation restrictions illustrated in (29). Although Bayesian analysis is still possible, the researcher will be forced to use numerical methods to characterize the mapping from the prior to the posterior given in (30). For example, for a given prior, numerical integration will have to be used to calculate the moments of the posterior distribution. My own judgment is that given current computer technology, formal Bayesian estimation procedures seem prohibitively expensive for most members of the class of dynamic models considered here. This is obviously not an objection to Bayesian methods in principle. However, I believe that the high cost attached to applying Bayesian methods correctly helps to explain why they have not yet been applied extensively to estimating rational expectations models.

A Second Model of the "Error Term"

More serious limitations on the domain of Bayesian techniques emerge if the researcher embraces a second model of the error term, which we now discuss. In the second model of the error term, it is assumed that the econometrician possesses only observations on subsets $\tilde{W}_t \subset W_t$ and $\tilde{D}_t \subset D_t$ of the information variables that private agents use to forecast future w_t's and D_{1t}'s.[43] It is assumed that these

[42] A point related to that raised in n. 29 is relevant here. Priors and posteriors that assign positive probability to points in regions for which zeroes of det $\delta(z)$ are less than $\sqrt{\beta}$ in modulus are inadmissible. This is because under such distributions, for some regions in the parameter space with positive probability, the dynamic optimum problems are not well posed. Taking this into account would substantially complicate the analysis since it would involve using mathematically less tractable distributions.

[43] This model of the error term was originally proposed by Shiller (1972) in a related but somewhat different context. The model was applied in the present context by Hansen and Sargent (1980*a*). Nerlove, Grether, and Carvalho (1979) also recommend Shiller's model of the error term.

subsets of information variables follow autoregressive processes $\tilde{\delta}_D(L)\tilde{D}_t = \tilde{V}_t^D$ and $\tilde{\delta}_w(L)\tilde{W}_t = \tilde{V}_t^w$, where $\tilde{\delta}_D(L)$ and $\tilde{\delta}_w(L)$ are polynomials in the lag operator of order $\tilde{r}_D$ and $\tilde{r}_w$, respectively. Then it turns out that the equilibrium law of motion for capital, (20), can be written in a form identical to (20), except that $W, D, \delta_D, \delta_w, H_D$, and H_w are to be replaced by the corresponding objects with tildes above them, *and* that there appears an additional random disturbance η_t on the right side of (20). The cross-equation restrictions (21) continue to characterize the objects with tildes over them.[44] The random variable η_t can be shown to be orthogonal to all of the current and lagged values of $\tilde{W}$ and $\tilde{D}$.[45] However, it turns out that η_t is in general serially correlated, with serial correlation properties that depend on the joint covariance properties of those variables in D_t and W_t that the econometrician does not have observations on. In the context of this setup, it is not even possible to write down the likelihood function without specifying details of the moments of information variables in D and W that are unobservable to the econometrician. It would seem attractive to adopt an estimation procedure that avoids the implicit theorizing about the stochastic properties of the unobserved D's and W's that an estimator using the likelihood function requires.

One such estimation strategy that exploits the orthogonality of η to $\tilde{D}$ and $\tilde{W}$, without requiring all of the added details required to write down a likelihood function, has been developed by Hansen (1979). The "generalized method of moments" estimators of Hansen have the advantage of delivering estimators of the free parameters whose desirable statistical properties do not depend on any arbitrary assumptions about the serial correlation properties of the η_t's.[46] These generalized method of moments estimators were invented precisely to handle situations in which the researcher is substantially more confident of the orthogonality conditions delivered by his theorizing than he is about the serial correlation properties of the error. These methods construct statistically consistent estimators, while avoiding the need to form the likelihood function. However, in acknowledging that he does not have enough information about the disturbances to construct the likelihood function, the researcher loses the ability to employ Bayesian methods, since knowledge of the likelihood function is essential for using Bayes's law as in (30).

[44] See Hansen and Sargent 1980*a*.
[45] Ibid.
[46] Under regularity conditions provided by Hansen (1979), the estimators of the underlying parameters are shown to be consistent and most efficient within a restricted class of estimators. Hansen's discussion of the conditions for consistency, which also has implications for the conditions for consistency of maximum likelihood estimators, is at this date the key reference on issues of statistical consistency in linear rational expectations models.

Interrelated Industries

In the Lucas-Prescott model of a single industry, state variables which help the firm predict the future prices of inputs appear in the representative firm's decision rule. The laws of motion of these input prices have been taken as given from outside the model. In actuality, the prices of these inputs are usually thought to be determined by trades in another market, one source of demand for which stems from the industry being modeled by Lucas and Prescott. If this other market is modeled explicitly, nontrivial modifications also occur in the analysis of the original industry. Thus, consider the example of a corn-hog model in which part of the output of one industry, corn, is an input into the production of the other industry, hogs. If the technology is such that hog producers have an incentive to forecast future corn prices, it follows that state variables that appear in the laws of motion for the total output and price of corn will also appear in the decision rule of the representative hog producer. Also, because the corn producer has an incentive to forecast the price of corn, which depends partly on the demand for corn from hog producers, the state variables that appear in the laws of motion for total hog output and the price of hogs will appear in the optimal decision rule of the representative corn producer. Hence, each industry inherits the state variables of the other. Furthermore, the equilibria in the two industries must be defined jointly, since the laws of motion for endogenous market-wide state variables are simultaneously determined for the two markets. There continues to be a fictitious social planning problem that a rational expectations equilibrium solves, one that is a version of an interrelated factor demand problem in which the planner is jointly optimizing the performance of both industries.[47]

These remarks indicate that the analyst will often face a hard practical decision about which dynamics he takes as given from outside the model. The internal logic of this class of models tends to propel the analyst toward a general equilibrium formulation in which the laws of motion characterizing each distinct industry are determined simultaneously with the laws of motion for all industries with which it buys and sells. In any given application, the researcher will have to choose what laws of motion he takes as given from outside the model, for the purposes of the analysis at hand.

Conclusions

Remaking dynamic econometric practice so that it is consistent with the principle that agents' constraints influence their behavior is a task

[47] These claims are proved for a particular version of a "corn-hog model" in some unpublished notes (Sargent 1980*a*).

that is far from finished. Further, properly allowing for the implications of the principle will surely require abandoning many presently received ways of interpreting data. A variety of setups can be imagined that are consistent with the principle. For example, a variety of variations of the setup of this paper can be imagined in which agents optimize but have smaller information sets than have been attributed to them here. Also, information discrepancies across classes of agents can be assumed. In many such cases, endogenous variables such as prices will play an important role in conveying information to agents. In models with dynamics as complicated as those of our examples, these variations introduce substantial analytical difficulties. To date there is very little work which investigates the econometric implications of such complications to setups like ours. With or without these complications, building a dynamic econometrics that is consistent with our simple principle from economic theory is a challenging task. It is sure to require substantial changes in the ways that applied economists interpret economic time series.

References

Almon, Shirley. "The Distributed Lag between Capital Appropriations and Expenditures." *Econometrica* 33 (January 1965): 178–96.

Anderson, Brian D. O., and Moore, John B. *Optimal Filtering.* Englewood Cliffs, N.J.: Prentice-Hall, 1979.

Anderson, Paul A. "Rational Expectations Forecasts from Nonrational Models." *J. Monetary Econ.* 5 (January 1979): 67–80.

Arzac, E. R., and Wilkinson, M. "Stabilization Policies for United States Feed Grain and Livestock Markets." *J. Econ. Dynamics and Control* 1 (February 1979): 39–58.

Bertsekas, Dimitri P. *Dynamic Programming and Stochastic Control.* New York: Academic Press, 1976.

Blackwell, David. "Discounted Dynamic Programming." *Ann. Math. Statis.* 36 (February 1965): 226–35.

Blanco, Herminio. "Investment under Uncertainty: An Empirical Analysis." Ph.D. dissertation, Univ. Chicago, 1978.

Cagan, Phillip. "The Monetary Dynamics of Hyperinflation." In *Studies in the Quantity Theory of Money*, edited by Milton Friedman. Chicago: Univ. Chicago Press, 1956.

Chow, Gregory C. "Multiperiod Predictions from Stochastic Difference Equations by Bayesian Methods." *Econometrica* 41 (January 1973): 109–18.

Craine, Roger. "Investment, Adjustment Costs, and Uncertainty." *Internat. Econ. Rev.* 16 (October 1975): 648–61.

Crawford, Robert G. "An Empirical Investigation of a Dynamic Model of Labor Turnover in U.S. Manufacturing Industries." Ph.D. dissertation, Carnegie-Mellon Univ., 1975.

Fischer, Stanley. "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule." *J.P.E.* 85, no. 1 (February 1977): 191–205.

Fisher, Franklin M. *The Identification Problem in Econometrics.* New York: McGraw-Hill, 1966.

Friedman, Benjamin. "Discussion." In *After the Phillips Curve: Persistence of High Inflation and High Unemployment.* Federal Reserve Bank of Boston Conference Vol. no. 19. Boston: Federal Reserve Bank, 1978.

———. "Optimal Expectations and the Extreme Information Assumptions of 'Rational Expectations' Macromodels." *J. Monetary Econ.* 5 (January 1979): 23–41.

Friedman, Milton. "The Methodology of Positive Economics." In *Essays in Positive Economics.* Chicago: Univ. Chicago Press, 1953.

———. *A Theory of the Consumption Function.* Princeton, N.J.: Princeton Univ. Press (for Nat. Bur. Econ. Res.), 1957.

Futia, Carl. "Rational Expectations in Speculative Markets." Unpublished paper, Bell Telephone Laboratories, 1979.

Geweke, John. "Wage and Price Dynamics in U.S. Manufacturing." In *New Methods in Business Cycle Research: Proceedings from a Conference*, edited by Christopher A. Sims. Minneapolis: Federal Reserve Bank, 1977.

Gordon, Donald F., and Hynes, Allan G. "On the Theory of Price Dynamics." In *Microeconomic Foundations of Employment and Inflation Theory*, by Edmund S. Phelps et al. New York: Norton, 1970.

Granger, C. W. J. "Investigating Causal Relations by Econometric Models and Cross-spectral Methods." *Econometrica* 37 (July 1969): 424–38.

Hall, Robert E. "The Macroeconomic Impact of Changes in Income Taxes in the Short and Medium Runs." *J.P.E.* 85, no. 2, pt. 2 (April 1978): S71–S85.

Hansen, Lars P. "Large Sample Properties of Generalized Method of Moment Estimators." Mimeographed. Pittsburgh: Carnegie-Mellon Univ., 1979.

Hansen, Lars P., and Sargent, Thomas J. "A Note on Wiener-Kolmogorov Prediction Formulas for Rational Expectations Models." Mimeographed. Pittsburgh: Carnegie-Mellon Univ., 1979.

———. "Formulating and Estimating Dynamic Linear Rational Expectations Models." *J. Econ. Dynamics and Control* 2, no. 1 (1980): 7–46. (*a*)

———. "Linear Rational Expectations Models for Dynamically Interrelated Variables." In *Rational Expectations and Econometric Practice*, edited by Robert E. Lucas, Jr., and Thomas J. Sargent. Minneapolis: Univ. Minnesota Press, 1980. (*b*)

Holt, Charles C.; Modigliani, Franco; Muth, John F.; and Simon, Herbert A. *Planning Production, Inventories and Work Force.* Englewood Cliffs, N.J.: Prentice-Hall, 1960.

Huntzinger, R. La Var. "Market Analysis with Rational Expectations: Theory and Estimation." *J. Econometrics* 10 (June 1979): 127–45.

Jorgenson, Dale W. "Rational Distributed Lag Functions." *Econometrica* 34 (January 1966): 135–49.

Kareken, John A.; Muench, Thomas; and Wallace, Neil. "Optimal Open Market Strategy: The Use of Information Variables." *A.E.R.* 63 (March 1973): 156–72.

Kennan, John. "The Estimation of Partial Adjustment Models with Rational Expectations." *Econometrica* 47 (November 1979): 1441–55.

Koyck, Leendert M. *Distributed Lags and Investment Analysis.* Amsterdam: North-Holland, 1954.

Kushner, Harold J. *Introduction to Stochastic Control.* New York: Holt, Rinehart, & Winston, 1971.

Kwakernaak, Huibert, and Sivan, Raphael. *Linear Optimal Control Systems.* New York: Wiley, 1972.

Kydland, Finn E., and Prescott, Edward C. "Rules Rather than Discretion: The Inconsistency of Optimal Plans." *J.P.E.* 85, no. 3 (June 1977): 473–91.

Leamer, Edward E. "A Class of Informative Priors and Distributed Lag Analysis." *Econometrica* 40 (November 1972): 1059–81.

———. *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* New York: Wiley, 1978.

Lucas, Robert E., Jr. "Expectations and the Neutrality of Money." *J. Econ. Theory* 4 (April 1972): 103–24. (*a*)

———. "Econometric Testing of the Natural Rate Hypothesis." In *The Econometrics of Price Determination: Conference, October 30–31, 1970, Washington, D.C.*, edited by Otto Eckstein. Washington: Board of Governors of the Federal Reserve System, 1972. (*b*)

———. "An Equilibrium Model of the Business Cycle." *J.P.E.* 83, no. 6 (December 1975): 1113–44.

———. "Econometric Policy Evaluation: A Critique." In *The Phillips Curve and Labor Markets*, edited by Karl Brunner and Allan H. Meltzer. Carnegie-Rochester Conferences on Public Policy, vol. 1. Amsterdam: North-Holland, 1976.

———. "Asset Prices in an Exchange Economy." *Econometrica* 46 (November 1978): 1429–45.

Lucas, Robert E., Jr., and Prescott, Edward C. "Investment under Uncertainty." *Econometrica* 39 (September 1971): 659–81.

Lucas, Robert E., Jr., and Sargent, Thomas J. "After Keynesian Macroeconomics." In *After the Phillips Curve: Persistence of High Inflation and High Unemployment.* Federal Reserve Bank of Boston Conference Vol. no. 19. Boston: Federal Reserve Bank, 1978.

———. "Rational Expectations and Econometric Practice" (introductory essay). In *Rational Expectations and Econometric Practice*, edited by Robert E. Lucas, Jr., and Thomas J. Sargent. Minneapolis: Univ. Minnesota Press, 1980.

Marschak, Jacob. "Econometric Measurements for Policy and Prediction." In *Studies in Econometric Method*, edited by William C. Hood and Tjalling C. Koopmans. Cowles Foundation Monograph no. 14. New Haven, Conn.: Yale Univ. Press, 1953.

Meese, Richard. "Dynamic Factor Demand Schedules for Labor and Capital under Rational Expectations." Unpublished paper. Washington: Board of Governors of the Federal Reserve System, 1979.

Merton, Robert C. "Optimum Consumption and Portfolio Rules in a Continuous-Time Model." *J. Econ. Theory* 3 (December 1971): 373–413.

Mishkin, Frederic S. "Simulation Methodology in Macroeconomics: An Innovation Technique." *J.P.E.* 87, no. 4 (August 1979): 816–36.

Modigliani, Franco. "The Monetarist Controversy or, Should We Forsake Stabilization Policies?" *A.E.R.* 67 (March 1977): 1–19.

Mosca, Edoardo, and Zappa, Giovanni. "Consistency Conditions for the Asymptotic Innovations Representation and an Equivalent Inverse Regulation Problem." *IEEE Transactions on Automatic Control* AC-24 (June 1979): 501–3.

Muth, John F. "Optimal Properties of Exponentially Weighted Forecasts." *J. American Statis. Assoc.* 55 (June 1960): 299–306.

———. "Rational Expectations and the Theory of Price Movements." *Econometrica* 29 (July 1961): 315–35.

Nerlove, Marc. "Distributed Lags and Unobserved Components in Economic Time Series." In *Ten Economic Studies in the Tradition of Irving Fisher*, by William Fellner et al. New York: Wiley, 1967.

Nerlove, Marc; Grether, David M.; and Carvalho, José L. *Analysis of Economic Time Series: A Synthesis.* New York: Academic Press, 1979.

Phelps, Edmund S., and Taylor, John B. "Stabilizing Powers of Monetary Policy under Rational Expectations." *J.P.E.* 85, no. 1 (February 1977): 163–90.

Prescott, Edward C., and Mehra, Rajnish. "Recursive Competitive Equilibrium: The Case of Homogeneous Households." *Econometrica* 48, no. 6 (September 1980): 1365–80.

Sargent, Thomas J. "A Note on the 'Accelerationist' Controversy." *J. Money, Credit and Banking* 3 (August 1971): 721–25.

———. "The Demand for Money during Hyperinflations under Rational Expectations. I." *Internat. Econ. Rev.* 18 (February 1977): 59–82.

———. "Estimation of Dynamic Labor Demand Schedules under Rational Expectations." *J.P.E.* 86, no. 6 (December 1978): 1009–44.

———. *Macroeconomic Theory.* New York: Academic Press, 1979.

———. "Lecture Notes on Optimal Control and Filtering." Mimeographed. Minneapolis: Univ. Minnesota, 1980. (*a*)

———. "Tobin's q and the Rate of Investment in General Equilibrium." In *On the State of Macro-Economics*, Carnegie-Rochester Conference Series, vol. 12, edited by Karl Brunner and Allan H. Meltzer. Amsterdam: North-Holland, 1980. (*b*)

Sargent, Thomas J., and Wallace, Neil. " 'Rational' Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule." *J.P.E.* 83, no. 2 (April 1975): 241–54.

Shiller, Robert J. "Rational Expectations and the Structure of Interest Rates." Ph.D. dissertation, Massachusetts Inst. Tech., 1972.

———. "A Distributed Lag Estimator Derived from Smoothness Priors." *Econometrica* 41 (July 1973): 775–88.

Sims, Christopher A. "Money, Income, and Causality." *A.E.R.* 62 (September 1972): 540–52.

———. "Macroeconomics and Reality." *Econometrica* 48 (January 1980): 1–48.

Taylor, John B. "Estimation and Control of a Macroeconomic Model with Rational Expectations." *Econometrica* 47 (September 1979): 1267–86.

———. "Output and Price Stability: An International Comparison." *J. Econ. Dynamics and Control* 2, no. 1 (February 1980): 109–32.

Telser, Lester G., and Graves, Robert L. *Functional Analysis in Mathematical Economics: Optimization over Infinite Horizons.* Chicago: Univ. Chicago Press, 1971.

Zellner, Arnold. *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley, 1971.

Activist Monetary Policy under Rational Expectations

Peter Howitt

University of Western Ontario

The purpose of this paper is to argue that the pursuit of an activist monetary policy may make economic sense even when people's expectations are formed rationally. The paper presents a simple model in which (1) prices are costly to adjust, (2) there is uncertainty concerning the parameters affecting aggregate demand, and (3) there are positive costs of gathering and processing information. By reference to this model it can be shown that an activist monetary policy may or may not be useful in offsetting aggregate disturbances, depending upon the extent of information costs and of parameter uncertainty.

I. Introduction

By an activist monetary policy I mean a policy whereby the monetary authority permits the money supply to react systematically to information concerning aggregate disturbances. The proposition commonly known as the Sargent-Wallace (hereafter SW) proposition[1] asserts that an activist monetary policy cannot succeed in offsetting aggregate disturbances when people's expectations are formed rationally unless the monetary authority possesses information un-

An earlier version of this paper was presented at the meetings of the American Economic Association, New York, December 1977. I am indebted to Russell Boyer, Stanley Fischer, Joel Fried, Herschel Grossman, Robbie Jones, David Laidler, Robert E. Lucas, Jr., Michael Parkin, Arthur Robson, and an anonymous referee for helpful comments and criticisms, and to the Social Sciences and Humanities Research Council of Canada for financial support.

[1] See Sargent and Wallace 1975. The proposition is derived in a more general context by Barro (1976).

[*Journal of Political Economy*, 1981, vol. 89, no. 2]

available to private agents. It asserts that in the absence of such superior information systematic variations in the money supply will be anticipated and thus neutralized even in the short run.

The SW proposition has been criticized by previous writers. Taylor (1975) pointed out that it will take private agents time before they learn the exact nature of the monetary policy.[2] During this transition period the SW proposition is invalid. Poole (1976), Fischer (1977), and Phelps and Taylor (1977) have pointed out that even if private agents possess the same information as the monetary authority they may not be able to act upon it immediately because of contractual precommitments, in which case the SW proposition is invalid. It has also been widely acknowledged that if the arguments underlying the SW proposition were valid it would be difficult to explain the observed persistence of movements in output and employment.[3]

The present paper focuses on another objection to the SW proposition: namely, that it fails to take into account (1) the costs of gathering and processing information and (2) uncertainty concerning the structure of the economic system. People may choose not to collect and process the monetary authority's information concerning aggregate disturbances when it is costly to do so.[4] Indeed, everyday observation suggests that most people pay no attention to such information. In such circumstances people may still be forming their expectations rationally, in the sense that they are making efficient use of (costly) information, yet they are not in a position to anticipate and neutralize the effects of an activist monetary policy. In order for the SW proposition to be true the monetary authority's information must not only be available to private agents it must also be economically usable by private agents.

This qualification to the SW proposition has been noted by Barro (1976), who argued nevertheless that the monetary authority could do just as well by disseminating its information as it could by varying the money supply. However, this paper attempts to show that such a disseminating policy may be inferior to an activist policy when there is uncertainty concerning the values of parameters in the economic system because (1) people may not have enough incentive to process information, and (2) even if people do process the information there is no guarantee that this will produce the same outcome as would a properly designed monetary policy. A demonstration of this argu-

[2] A similar point was made by Benjamin Friedman (1979), who showed the similarity of rational expectations to error learning during the learning process.

[3] See, however, Lucas 1975 and Sargent 1976.

[4] This point has been stressed by Rutledge (1974), Frenkel (1975), Feige and Pearce (1976), Laidler (1978), Shiller (1978), and Friedman (1979).

ment can be made precise only in terms of a formal model, but it can be sketched in the following terms.

First, even if the information concerning aggregate disturbances is provided to people they must still incur the cost of processing it if they are to use it. But they may not find it economical to incur that cost if they think that not many other people will be using the information, even when its social value exceeds the processing cost. Consider, for example, an agent who is deciding what price to charge for the single good that he sells. Suppose he is told that the demand for money has doubled. If he thinks that all other agents are using this information he can calculate that his best course of action will be to cut his price in half, along with everyone else. The information may be quite valuable to him and to society because failure to conform in reducing his price would leave him charging a suboptimal relative price. But if he thinks that no other agent is using the information his best course of action will depend on the effects of that change in the demand for money on the demand for his product. If this effect is difficult to estimate accurately then the information may have little value to him. A convention like universal indexing that leads people to believe that everyone else will be using the information as well may be necessary for the private value of such information to be as great as its social value. Everyday observation again suggests that such a convention rarely exists.

Second, in a world without complete Arrow-Debreu contingent-claim markets there is no guarantee that even an otherwise frictionless competitive equilibrium will be efficient with respect to the information possessed by private agents. It is possible that the monetary authority is in a position to make better use of aggregate information than are private agents, even if the above-mentioned indexing convention could be established. In the example of the previous paragraph, Brainard's (1967) analysis suggests that if the values of the parameters of the economic system are not known with certainty then the optimal reaction of the monetary authority is to allow the money supply to increase, but by less than 100 percent. But if the information is disseminated prices may indeed fall in half, which would be equivalent to a 100 percent increase in the money supply. As in the theory of the second best, full information results in the best equilibrium, but when information is incomplete then more information does not necessarily result in a better equilibrium.[5]

[5] This result is similar to those of Hirshleifer (1971) and Grossman and Stiglitz (1976), who argued that the equilibrium allocations with more information may be Pareto inferior to those with less.

The remainder of the paper attempts to make these arguments more precise by developing a simple model of a monetary economy in which prices are costly to adjust. Section II describes the basic features of the model; Section III analyzes the private decision whether or not to use information; Section IV characterizes the optimal policy of the monetary authority showing how an activist policy may be justified by the first of the above arguments; Section V considers a variant of the model in which the second of the above arguments may be used to justify an activist policy; and Section VI contains some concluding remarks.

II. The Basic Model

Preliminaries

This section describes a simple hypothetical economy possessing the features described above. In this economy there are two tradable objects: money and a good. There are also two classes of economic actors: a monetary authority, whose job it is to determine how much money will exist and to distribute the money, and a set of private agents (referred to simply as "agents"), who produce and consume the good. In the interest of simplicity we suppose that each agent possesses identical tastes, technology, and initial endowments. (Thus we are making no distinction between firms and households.) In order to establish a basis for exchange among these identical agents we assume that each of them has an absolute aversion to consuming units of the good that he himself has produced. Thus he is induced to sell all of these units and to purchase for consumption units of the good that are produced by others.

To keep the demand and supply functions simple we suppose that each agent evaluates his demands according to a Stone-Geary utility function and produces subject to a quadratic cost function. That is, each agent's utility during a period depends upon $\hat{q}$, the amount of the good consumed during the period, $\hat{m}$, the real value of money balances held at the end of the period, and q, the amount of the good produced during the period, according to the particular function

$$U = K\hat{q}^{a}(\hat{m} + x)^{1-a} - (k/2)q^{2} \tag{1}$$

where k is a positive, nonstochastic parameter equal to the slope of the agent's marginal cost schedule, a and x are independent stochastic parameters with $0 < a < 1$, and K is a stochastic parameter defined as $K \equiv a^{-a}(1 - a)^{-(1-a)}$. For reasons that will become apparent later we shall refer to the parameter x as "the level of aggregate demand."

In order for purely monetary changes potentially to have real effects even in the short run we must not assume that there exists a central Walrasian auctioneer who is able each period to compute a single market-clearing price at which all trades are to occur. The actual transaction prices must be determined before the agents have collected enough information for them to be able to compute this Walrasian equilibrium price. We accomplish this by assuming that each agent is responsible for setting his own selling price. He must set the price before learning exactly how much he will be able to sell, and all trades must occur at these posted prices whether or not markets are clearing in the usual sense.[6] Assume also that the quantities transacted at these prices always equal the quantities demanded. In other words, when markets are not clearing the sellers bear all the resulting quantity rationing.[7]

At the time of the price-setting decision each agent's expected utility will depend negatively on the variance of his error in forecasting demand, because he must produce subject to increasing marginal cost. Thus he will have an incentive before setting his price to gather and process information concerning the demand function facing him so as to reduce the variance of this forecast error. The decision on what information to use at this stage will be the main focus of the following analysis.

When determining the value of M, the nominal per capita stock of money, the monetary authority is assumed to be unaware of the exact state of aggregate demand. Suppose, however, that he is able to acquire partial information by observing a variable, s, that is correlated with aggregate demand. In particular, assume that

$$x = cs \tag{2}$$

[6] This setup may be contrasted with Lucas's (1972), in which the agents are randomly divided into subgroups, each with a miniature Walrasian auctioneer, and with no communication between groups during the market period. Both setups require agents to make trading commitments before being completely informed of other agents' dispositions to trade. Lucas's approach has the advantage that market-clearing models have been more thoroughly analyzed in the literature than price-setting models. Ours has the advantage of replacing the Walrasian auctioneer with a scheme that bears some resemblance to the way prices actually get determined in many real-world markets. Whether our results could be transposed to a suitably constructed "market-clearing" model is an open question.

[7] This rule also appears to be compatible with the way transaction quantities are determined in many real-world markets, especially those organized by specialist traders—wholesalers, retailers, etc.—who post prices and hold inventories which they allow to act as buffer stocks against fluctuations in demand. By absorbing the nonprice rationing they provide a service to customers, for which they are rewarded by a spread between their buying and selling prices. In effect we are assuming that each agent is a combination consumer-producer-trader. (However, we are ignoring the potentially important role of inventories by assuming that the good is not storable.)

where c is a stochastic parameter. We shall refer to s as the "indicator" of monetary policy. The monetary authority is assumed to obey the following rule:[8]

$$M = \overline{M}(1 - gs) \tag{3}$$

where $\overline{M}$ (> 0) is the average per capita stock of money and g is the "policy reaction coefficient." Thus the monetary authority follows a "feedback" rule of reacting to observable information concerning the level of aggregate demand, and its two decision variables are $\overline{M}$ and g. The SW proposition asserts that the stochastic behavior of aggregate output in the economy will be independent of the monetary authority's choice of g.

The Sequence of Activities

During each period activities proceed according to the following sequence. First, there is a stage in which the monetary authority decides on his policy. That is, he chooses values for g and $\overline{M}$. When making this decision he is assumed to possess an accurate model of the economy, including exact knowledge of all deterministic variables, parameters, and functions, and the probability distributions of all stochastic variables. He does not, however, know the exact value that will be assumed during the period by any stochastic variable. As soon as he makes his decision all of his information is conveyed to each private agent, including the chosen values of g and $\overline{M}$. Thus there is no scope for the monetary authority to fool private agents, even in the short run, about the nature of his policy, or to make this policy decision using information that is not available to private agents.

Second, there is a stage in which private agents make their monitoring decision. That is, each one must choose whether to pay the cost of learning the exact value of s before setting his price, in which case he will be able to reduce the variance of his forecast error, or to wait until after his price has been set, when the value of s will be revealed to him at no cost. In the interest of simplicity we assume that the indicator is the only stochastic variable whose value is potentially observable before the price-setting decision is made.

Third, there is a stage in which the exact value of the indicator s is determined. Any agent who has decided to monitor observes s and pays the cost at this stage, as does the monetary authority (unless he has set $g = 0$). Note that once again we are not giving the monetary authority any informational advantage. At this stage, before he

[8] A purely random, unpredictable component could be added to (3) but, as Barro (1976) has also shown, the optimal policy, if feasible, would be to set its variance to zero.

chooses his price anyone can choose to be as well informed as the monetary authority.

The fourth stage is the one in which the values of all prices and of the money supply are decided. The money supply is determined automatically by equation (3), and the private agents choose their prices in such a way as to maximize expected utility. If they have decided to monitor the indicator (to learn the value of s), then private agents can forecast the exact value of M by using equation (3) and can use this knowledge in making their price decision. Otherwise they must make their decisions on the basis of expectations concerning M, which they form using (3) and their knowledge of the probability distribution of s. When setting his own price no agent is able to observe the others' prices. However, he is able to tell whether or not the others are monitoring the indicator, and this will be enough in most cases to allow him to predict exactly what the others' prices will be.

The fifth stage is the one in which all uncertainty is resolved. In this stage the monetary authority distributes money to the private agents in equal amounts, and the exact values of a, x, K, c, and s are revealed to everyone. Whatever values of the utility parameters a, x, and K are determined in this stage, they are shared by all agents in common. Thus, with this information each agent is able (i) to calculate exactly how much his sales will be for the period and (ii) to compute his own demands for money and the good.

In the sixth and final stage the agents visit one another to execute their planned purchases, and each agent must produce to meet the demands placed upon him. Up until this stage each of the identical agents will have been led to make identical decisions. Thus every agent will be charging the same price. Assume that in such a situation the quantity demanded from each agent will be identical.[9]

To understand what happens in this sequence we must start at the end and work back. Consider an agent who, during the sixth stage, is going to be called upon to produce the quantity q. Suppose that all other agents have set the same price, P, and that he himself has set the price rP. Then during the fifth stage he will be aware of the value of q and of the exact values of all the parameters in his utility function (eq. 1). He will formulate his own demands, $\hat{q}$ and $\hat{m}$, so as to maximize this utility function subject to the budget constraint

$$\hat{m} + \hat{q} = m + rq \tag{4}$$

[9] Alternatively, we could suppose that they all face a stochastic demand function with independent, identical probability distributions, the randomness representing the indeterminacy in the distribution of sales among identical sellers charging identical prices. However, the addition of one more random factor would add nothing but notational complexity to the analysis.

where $m = M/P$ denotes his initial holding of real balances.[10] Thus his demand for the good will be

$$\hat{q} = a(m + rq + x), \tag{5}$$

and by substituting from (4) and (5) into (1) we can see that the level of utility that he will obtain is determined by the indirect utility function

$$V(m, q, x, r) = m + rq + x - (k/2)q^2. \tag{6}$$

During the fifth stage each agent will also be able to predict his sales, q, in the following way. First, suppose that every agent is charging the same price, P. Then every agent will sell the same quantity, q, and will formulate demands according to (5) with $r = 1$. But every agent will also realize that his sales, q, will be equal to the per capita demand, $\hat{q}$. Thus the values of q and $\hat{q}$ can be computed by substituting $q = \hat{q}$ into (5), with $r = 1$, to produce

$$q = b(m + x) \tag{7}$$

with

$$b = \frac{a}{1 - a} > 0. \tag{8}$$

Equation (7) indicates the amount that each agent would demand in stage 5, as well as the amount that he would expect to sell, under the assumption that $r = 1$. And indeed the equilibrium will be one in which $r = 1$. But to see how prices are chosen we must indicate how much each agent would expect to sell if $r \neq 1$. Let us suppose that this quantity is given by the expression

$$q(m, x, b, r) = \max\,[b(m + x) + n(1 - r), 0] \tag{9}$$

where $n > 0$ is a measure of the "conjectured competitiveness" in the economy. In other words, it measures the extent to which each agent supposes that he would lose sales by raising his price above the market price. The limiting case in which $n = +\infty$ is the one usually associated with the assumption of perfect competition.

[10] It has been argued that even in markets where demanders choose quantities at predetermined prices the demand decisions must take into account the future relationships between trading partners, as well as the usual objectives, in which case the quantities chosen may be on neither the demand nor the supply curve as usually conceived (Barro 1977). This consideration is probably important in markets where exchange is conducted on a highly personal basis because of the costs of switching trading partners, as in labor markets, markets for personal credit, and so forth. However, there are also many markets where exchange is quite impersonal and buyers may, because they can remain anonymous to the sellers, choose their quantities without taking this consideration into account. When we assume that demanders choose their quantities according to the usual sort of utility maximization we are implicitly assuming this sort of impersonal arrangement.

We now have introduced a total of seven random variables: b, c, s, x, a, K, and M. Let us suppose that the probability distributions of the first three of these are mutually independent and that

$$\begin{aligned} &E(b) = \beta > 0; E(c) = \gamma > 0; E(s) = 0 \\ &\text{var}(b) = \sigma_b^2 > 0; \text{var}(c) = \sigma_c^2 > 0; \text{var}(s) = \sigma_s^2 > 0. \end{aligned} \tag{10}$$

At the outset of the fourth stage each agent realizes that if he sets the price rP his utility will equal

$$\frac{\overline{M}}{P}(1 - gs) + cs + rq\left[\frac{\overline{M}}{P}(1 - gs), cs, b, r\right] - \left(\frac{k}{2}\right)q\left[\frac{\overline{M}}{P}(1 - gs), cs, b, r\right]^2. \tag{11}$$

(This is obtained by substituting from [2], [3], and [9] into [6].) He will know that the stochastic variables b, c, and s obey (10), and he may or may not know the exact value of s, depending on whether or not he has chosen to monitor s in the second stage. Suppose that in stage 2 he has made the same monitoring decision as every other agent. Then he will be able to calculate P, the price set by all other agents. He will also know the values of $\overline{M}$, g, and k. Thus he will choose a relative price so as to maximize the expected value of (11) conditional on the information set I, where $I = \{s\}$ if he has monitored and $I = \emptyset$ otherwise. The first-order condition of this maximization (assume that $q[m, x, b, r] > 0$ *w.pr.* 1) can be expressed as

$$E\{b[\overline{M}(1 - gs)/P + cs] + n(1 - r) \mid I\} = r\bar{q}_n \tag{12}$$

where

$$\bar{q}_n = n/(1 + nk). \tag{13}$$

Let the superscript i denote the information set, $i = 1$ denoting $I = \{s\}$ and $i = 2$ denoting $I = \emptyset$. Then the solution to (12) can be written as the relative-price function

$$r = r_n^i(P, \overline{M}, g, s), i = 1, 2. \tag{14}$$

In order to tell what price to set each agent must also calculate P, the price that the other agents each will set. To do this he calculates the value of P that makes the value of r in (14) equal to unity. That is, he expects the price to be one such that every agent (including himself) is induced to conform. This will also be the actual price set by all agents in stage 4. It can be expressed as

$$P = P_n^i(\overline{M}, g, s), i = 1, 2, \tag{15}$$

and, from (12), it must satisfy the condition

$$E(q \mid I) \equiv E\{b[\overline{M}(1 - gs)/P + cs] \mid I\} = \bar{q}_n. \tag{16}$$

In the limiting case of perfect competition, (16) states that the market price will be set so that the expected market demand (per capita) equals the competitive market supply (per capita), $\bar{q} \equiv \lim_{n\to\infty} \bar{q}_n = 1/k$. In other words, there is expected market clearing.[11]

Let $\bar{m}_n \equiv \bar{q}_n/\beta$. From (10), (15), and (16) the prices can be expressed as

$$P_n^1(\overline{M}, g, s) = \overline{M}(1 - gs)/(\bar{m}_n - \gamma s), \tag{17}$$

$$P_n^2(\overline{M}, g, s) = \overline{M}/\bar{m}_n. \tag{18}$$

Thus the quantity of output that will be produced per person in stage 6 can be expressed as a function of the information set, the reaction coefficient g, and the random variables b, c, and s by substituting from (2), (3), (17), and (18) in (7) to obtain

$$q_n^1 = b[\bar{m}_n + (c - \gamma)s], \tag{19}$$

$$q_n^2 = b[\bar{m}_n(1 - gs) + cs]. \tag{20}$$

The essence of the SW proposition is contained in (19). If agents monitor s then the behavior of the quantity q is independent of the reaction coefficient g. On the other hand, (20) shows that q is affected by g if no monitoring occurs.

From (10), (19), and (20) the mean and variance of output per person, as seen at the beginning of stage 3 (i.e., before the exact value of s has been ascertained), can be expressed as

$$E(q_n^1) = E(q_n^1 \mid s) = E(q_n^2) = \bar{q}_n \neq E(q_n^2 \mid s), \tag{21}$$

$$\text{var } q_n^1 = \sigma_b^2 \bar{m}_n^2 + \sigma_s^2(\beta^2 + \sigma_b^2)\sigma_c^2, \tag{22}$$

$$\text{var } q_n^2 = \sigma_b^2 \bar{m}_n^2 + \sigma_s^2(\beta^2 + \sigma_b^2)[\sigma_c^2 + (\gamma - g\bar{m}_n)^2]. \tag{23}$$

From (6) and (17)–(21) the utility that each agent expects at the end of stage 2 may be expressed as

$$V_n^i(g) = EV(M/P_n^i, q_n^i, x, 1) = \bar{m}_n + \bar{q}_n - (k/2)\bar{q}_n^2 - (k/2) \text{ var } q_n^i;\ i = 1, 2. \tag{24}$$

It follows from (22)–(24) that the gain to society from everyone monitoring the indicator (i.e., $V_n^1[g] - V_n^2[g]$) arises solely from the consequent reduction in the variance of output. This reduction in variance yields a welfare gain because the marginal utility of more production is (by assumption) a constant but the marginal cost is increasing, so that total utility is a concave function of the level of output.

[11] An almost identical condition was *assumed* to hold by Fischer (1977, pp. 195–97) and by Phelps and Taylor (1977, pp. 166–70).

III. The Monitoring Decision

Equations (17)–(24) show how each agent's price, output, and expected utility will depend on the monitoring decision made in stage 2 and the reaction coefficient chosen in stage 1. This section analyzes the monitoring decision.

Let z_m (> 0) denote the cost, in utils, of monitoring the indicator. This cost will generally equal $z_c + z_p$ where z_c is the cost of collecting the information (i.e., of determining the exact value of s) and z_p is the cost of processing the information (i.e., of computing and posting the optimal price given the value of s). Each agent will choose to monitor if the gain to him in expected utility is at least as great as z_m.

The size of this gain will depend on whether or not the other agents are monitoring. Consider first the case in which none of the other agents is monitoring. In this case the agent will derive the expected utility $V_n^2(g)$, as determined by (24), if he too chooses not to monitor. But if he chooses to monitor, the market price will continue to be P_n^2, as in (18), and he can now profit from knowing the exact value of the indicator before setting his own price. Because he can infer, as before, the exact value of P before determining his relative price, that relative price will be the one that satisfies the first-order condition (12) with $I = \{s\}$ and $P = P_n^2(\overline{M}, g, s)$. That is,

$$r_n^1(P_n^2, \overline{M}, g, s) = 1 + s\beta(\gamma - g\bar{m}_n)/(\bar{q}_n + n). \tag{25}$$

Let $V_n^3(g)$ denote the expected utility of the agent who monitors when no one else is monitoring. This can be obtained by substituting from (25) into (6) and using $q = q_n^3 \equiv q_n^2 + n(1 - r)$ with q_n^2 given by (20). As a result of this calculation it can be seen that the gain in expected utility from deciding to monitor when others are not monitoring is

$$V_n^3(g) - V_n^2(g) = \phi(n)\sigma_s^2\beta^2(\gamma - g\bar{m}_n)^2 \tag{26}$$

where $\phi(n) = (1 + nk)^2/2n(2 + nk)$.[12] Note that

$$\phi(n) > (k/2) \text{ and } \phi(n) \to (k/2) \text{ as } n \to \infty. \tag{27}$$

The nature of this gain is easy to see. The expected output of the agent is not affected by the decision to monitor but its variance is, and, as we have argued, a reduction in variance increases expected utility. Specifically, whenever the level of aggregate demand is high ($s > 0$) the monitoring agent will, according to (25), raise his price, thereby damping the increase in his sales (assuming that $g < \gamma\bar{m}_n^{-1}$). Likewise, when $s < 0$ the monitoring agent will bolster his sales by reducing r. Not only does this reduce variability it also adds to the agent's ex-

[12] This result is also derived in the Appendix.

pected revenue. In the limiting case of perfect competition, (25) implies that these changes in r will be imperceptible. In this case the agent is able to adjust the mean position of his random sales schedule whenever a nonzero value of s is observed so that the expected value of his sales *conditional upon* s is equal to $\bar{q}$,[13] and the entire gain to the agent from deciding to monitor arises from the resulting reduction in the variance of sales. In particular, (26) and (27) imply that the gain is equal to $(k/2)$ times this reduction in variance.[14]

Next, consider the case in which all of the other firms are monitoring. In this case the agent will derive the expected utility $V_n^1(g)$ as determined by (24) if he too chooses to monitor. But if he chooses not to monitor then the price will continue to be P_n^1 as in (17), and the agent will lose by not being able to foresee, before setting his own price, the variations in s and in P_n^1 that his competitors are taking into account. The nonmonitoring agent will choose a price $\hat{P}$ so as to maximize $EV[M/P_n^1, q_n^1 + n(1 - \hat{P}/P_n^1), x, \hat{P}/P_n^1]$. Let $V_n^4(g)$ denote the expected utility of the agent who does not monitor when others are monitoring. Then, as the Appendix demonstrates, the gain in expected utility from deciding to monitor when others are monitoring is

$$V_n^1(g) - V_n^4(g) = \psi(n)\frac{\operatorname{var} e_n}{1 + \operatorname{var} e_n} \tag{28}$$

where $e_n = P_n^1(\overline{M};\, g,\, s)^{-1}[EP_n^1(\overline{M};\, g,\, s)^{-1}]^{-1}$ is the reciprocal of the market price, expressed as a ratio to its mean value, and where $\psi(n) = n + n^2k/2$. Note that

$$\psi(n) \to \infty \text{ as } n \to \infty. \tag{29}$$

If everyone is to monitor then the gain from deciding to monitor when everyone else is monitoring must be at least as great as the monitoring cost; that is,

$$V_n^1(g) - V_n^4(g) \geqq z_m. \tag{30}$$

Likewise, if no one is to monitor, it is necessary that

$$V_n^3(g) - V_n^2(g) \leqq z_m. \tag{31}$$

From (26) and (28) it follows that for large enough n at least one of these conditions must be satisfied. But it also follows that both of them

[13] The limiting case of perfect competition is denoted by the absence of the subscript n on the variables $\bar{q}_n$, $\bar{m}_n$, V_n^1, $\hat{g}_n$, etc.

[14] Notice that we are not allowing a monitoring agent to gain by selling his information. Allowing a market for information would complicate the analysis without affecting our main result. For in the case that we focus on below where an activist policy dominates all others the private value of information is less than the cost of processing it. In this case no one would willingly pay the monitoring agent anything for his information.

may be satisfied. In other words, convention appears to play an important role in the monitoring decision. It may be that everyone will monitor if they all expect everyone else to, but no one will monitor if they all expect no one else to.[15] I shall assume that monitoring will occur if and only if (31) is violated—in other words, that the convention of monitoring will develop only when the convention of non-monitoring is unsustainable.[16] Notice, however, that when the reaction coefficient assumes the value

$$\hat{g}_n = \gamma/\bar{m}_n \tag{32}$$

then the convention of monitoring could never be sustained as long as $z_m > 0$, because from (17), (26), and (28),

$$V_n^1(\hat{g}_n) - V_n^4(\hat{g}_n) = V_n^3(\hat{g}_n) - V_n^2(\hat{g}_n) = 0. \tag{33}$$

In order to simplify the analysis further, let us assume the limiting case of perfect competition. Then, from (2), (26), (27), (31), and (32) the necessary and sufficient condition for monitoring to occur is

$$(k/2)\lambda_c\sigma_x^2\beta^2(1 - g/\hat{g})^2 > z_m \tag{34}$$

where $\lambda_c = \gamma^2/(\gamma^2 + \sigma_c^2)$ is a measure of the accuracy of the indicator, being the square of the coefficient of correlation between x and the conditional forecast γs of x based on the indicator, $\sigma_x^2 = \text{var } x$ is the variance in aggregate demand, and $\hat{g} = \gamma\beta k$, the limiting value of (32).

According to (34), the likelihood of monitoring occurring is larger (i) the larger is the slope, k, of the typical producer's marginal cost schedule, because the larger this slope the more the firm loses from variability in its demand that could be avoided by monitoring; (ii) the

[15] In such cases the private value of the information contained in s is larger when everyone else has the information than when no one else has it. This result may appear somewhat counterintuitive, but it arises from the following considerations. In stage 4 the agent who knows s may gain because this allows him to vary his price in anticipation of variations in (a) the variable x which is correlated with s, and (b) his competitors' price, which may be predicted once s is known. When n is large, then b will be the dominant factor if others are observing s, because variations in his competitors' absolute selling price imply variations in his own relative selling price, and thus large variations in his sales, if he is unable to anticipate them. But if no one else is observing s then b no longer applies because he does not need to observe s in order to anticipate his competitors' price. More generally, the result arises because the main use of the information is to allow the agent to coordinate his activities (specifically the activity of setting his nominal selling price) with the activities of other agents. As an analogy, the value of knowing that you should drive on the left in England would be much less if none of the other drivers knew it. This analogy also illustrates why multiple equilibria that must be selected by convention can arise under such circumstances.

[16] Thus we may say that an equilibrium with respect to any $(g, \bar{M}, n)$ consists of a monitoring decision $i \in \{1, 2\}$ and a price function $P_n^i(\bar{M}, g, s)$ such that (a) for any s the optimal relative price $r_n^i(P, \bar{M}, g, s)$ equals unity when $P = P_n^i(\bar{M}, g, s)$, and (b) the monitoring decision is $i = 1$ only if $V_n^1(g) - V_n^4(g) \geqq z_m$, and $i = 2$ if and only if $V_n^3(g) - V_n^2(g) \leqq z_m$.

more accurate is the indicator, because an inaccurate indicator conveys little useful information; (iii) the more variable is the level of aggregate demand, because there is little gain in monitoring a variable that is relatively constant; (iv) the larger is β, because there is little gain to monitoring a variable if its expected effect on demand is small; and (v) the further is the reaction coefficient from $\hat{g}$, because when $g = \hat{g}$ then, as can be inferred from comparing (22) with (23), the monetary authority is already accomplishing everything that private agents could accomplish by monitoring.

IV. Optimal Monetary Policy and the Incentive to Monitor

Three implications of this analysis are worth noting here. First, the SW proposition is valid for all values of g satisfying (34). For in this case monitoring will occur, and, as noted before, the behavior of output is unaffected by the value of g. According to (22) and (24) each agent's expected utility will also be unaffected by the value of g.

Second, whatever can be accomplished by monitoring can be accomplished just as well by having the monetary authority set $g = \hat{g}$. According to (34) no monitoring would occur, but according to (19), (20), and (24) the behavior of output and expected utility would be the same as if monitoring was occurring.

Third, as was argued in the introduction above, private agents may not have much incentive to monitor s even when the social gains from everyone deciding to monitor are large. That is, the private value of monitoring, $V^3(g) - V^2(g)$, may be small when the social value, $V^1(g) - V^2(g)$, is large. In particular, it follows from (22)–(24) and (26) that

$$[V^3(g) - V^2(g)] = \lambda_b[V^1(g) - V^2(g)], \tag{35}$$

where $\lambda_b = \beta^2/(\beta^2 + \sigma_b^2)$ is a measure of the predictability of the effect upon effective demand of changes in x, being the square of the coefficient of correlation between the actual effect bx and the predicted effect βx. If the value of the coefficient b is known with certainty, then the private and social values coincide. In the limiting case of pure uncertainty (when $[\beta/\sigma_b]^2 \to 0$) the private value is infinitesimal in comparison with the social value.

The reason for this divergence of private from social value is seen most easily in the case of a perfectly reliable indicator ($\sigma_c^2 = 0$) and a neutral monetary policy ($g = 0$). If everyone monitors then the price level will adjust to fluctuations in x so that the sum $m + x$ is constant and equal to $\bar{m}$, with output equal to $q^1 = b\bar{m}$. Thus everyone monitoring together will completely eliminate the effects of x upon output. An isolated agent monitoring when no one else is cannot stop $m + x = \bar{m} + x$ from fluctuating. He can, and will, through impercep-

tible changes in r alter the mean position of his demand curve so as to offset the predicted effect βx, but there will still remain an unpredicted effect $(b - \beta)x$. Specifically, by substituting (25) into (9) we see that his output will be $q^3 = b\bar{m} + (b - \beta)x$. Thus the social gain to monitoring comes from eliminating the total effect bx of changes in aggregate demand, whereas the private gain comes from eliminating only the predicted effect βx. The ratio λ_b of private to social gain is just the ratio of variances of the predicted and total effects.

Consider three different monetary policies. (i) Policy N is the neutral policy, $g = 0$, which is assumed to result in no administrative cost. (ii) Policy D is the disseminating policy suggested by Barro of setting $g = 0$ but also publicizing the value of s, which is assumed to result in the administrative cost z_d per person. (iii) Policy A is the activist policy of setting $g = \hat{g}$, resulting in the administrative cost z_a. Assume that the cost z_a is incurred whenever *any* nonzero value of g is chosen. Then policy A will dominate any other policy with $g \neq 0$, and we may restrict our attention to policies A, D, and N. Which policy is optimal will depend upon the relationships between the costs and benefits of monitoring and the administrative costs of the different policies. Let

$$V^s \equiv V^1(0) - V^2(0) = (k/2)\lambda_c \operatorname{var}(bx) \tag{36}$$

denote the social gain to monitoring and $V^p \equiv V^3(0) - V^2(0) = \lambda_b V^s$ denote the private gain to monitoring, when $g = 0$. Then according to (34) monitoring will never occur under policy A and will not occur under policy N if $\lambda_b V^s \leqq z_c + z_p$. Policy D is assumed to save the private agents any collecting costs, but if they are to use their knowledge of the value of s they must still incur the processing cost. Thus monitoring will not occur under policy D if $\lambda_b V^s \leqq z_p$. Table 1 depicts the cost per person of the different policies in the three situations indicated, where the cost is measured as $V^1(0)$ minus the expected utility resulting from the policy plus the monitoring and administrative costs. In each situation the optimal policy is the one with the least cost.

TABLE 1

COST PER PERSON OF MONETARY POLICY

	Policy		
Situation	N	D	A
I. $z_c + z_p < \lambda_b V^s$	$z_c + z_p$	$z_d + z_p$	z_a
II. $z_p < \lambda_b V^s \leqq z_c + z_p$	V^s	$z_d + z_p$	z_a
III. $\lambda_b V^s \leqq z_p$	V^s	$z_d + V^s$	z_a

NOTE.—N = neutral policy, D = disseminating policy, A = activist policy.

Note that if the indicator is unreliable enough (i.e., λ_c small enough) or if effective demand is stable enough (var $[bx]$ small enough), *ceteris paribus*, then a neutral policy will be optimal because both of these factors make V^s small. Both of these factors also figure prominently in Milton Friedman's (1953, 1968) case against activist policy. For the same reason, a small slope to marginal cost (small k) favors a neutral policy. Note also that in situation III the disseminating policy is dominated by the neutral policy.

For our purposes the most important implication to note about table 1 is that an activist policy will be uniquely optimal whenever $z_a < V^s \leqq z_p/\lambda_b$; that is, if (i) the indicator is reliable enough, effective demand is unstable enough, or marginal cost rises fast enough that V^s exceeds the cost of administering an activist policy; and (ii) the cost of processing the information contained in the indicator is large enough or the predictability of the effect of changes in the level of aggregate demand is small enough that $V^s \leqq z_p/\lambda_b$. These conditions do not require the monetary authority to possess any cost advantage. If λ_b is very small, they may hold even if the cost of administering an activist policy is much larger than the private cost of monitoring. If the conditions hold then an activist policy is optimal because despite the large social gain to monitoring no one will do it even if the cost of collecting information is saved. For even under a disseminating policy the user of the information contained in the indicator must incur a processing cost, and, as we have seen, he will be led to underestimate the gains from using the information.

V. A Variant of the Basic Model: Monetary Policy and the Efficiency of Monitoring[17]

One special feature of the basic model employed above is that the terms m and x enter into the aggregate demand function (5) with the same coefficient. This is why the optimal reaction coefficient $\hat{g}$ always makes the expected level of output (conditional on s) equal to its optimal value, $\bar{q}$ (see [20]). When m and x have different random coefficients then Brainard's (1967) analysis suggests that the optimal value of the reaction coefficient will generally be smaller than this. Such a case is easily constructed by replacing the utility function (1) with the variant

$$U = K(\hat{q} - x)^a \hat{m}^{1-a} - (k/2)q^2 + 2x, \tag{37}$$

in which case the indirect utility function is still given by (6), but the

[17] The rationalization for activist policy presented in this section is similar to the one presented by Fane (1977).

effective demand function changes from (7) to

$$q = bm + x. \tag{38}$$

As in (19) and (20), the value of output with or without monitoring can be expressed as

$$q_n^1 = b(\bar{m}_n - \gamma s/\beta) + cs \tag{39}$$

$$q_n^2 = b\bar{m}_n(1 - gs) + cs. \tag{40}$$

Once again the expected value of output is

$$E(q_n^1) = E(q_n^1 \mid s) = E(q_n^2) = \bar{q}_n \tag{41}$$

and the variance of output is

$$\operatorname{var} q_n^1 = \sigma_b^2\bar{m}_n^2 + \sigma_s^2[(\sigma_c^2 + \gamma^2\sigma_b^2)/\beta^2] \tag{42}$$

$$\operatorname{var} q_n^2 = \sigma_b^2\bar{m}_n^2 + \sigma_s^2[\sigma_c^2 + (\gamma - g\bar{m}_n\beta)^2 + (g\bar{m}_n)^2\sigma_b^2]. \tag{43}$$

Let us again restrict our attention to the limiting case of perfect competition. As before, the optimal value g' of the reaction coefficient will be the one that minimizes var (q^2). That is,

$$g' = \lambda_b\gamma/\bar{q}. \tag{44}$$

If the activist policy $g = g'$ is pursued and no monitoring occurs, then

$$E(q^2 \mid s) = \bar{q} + (1 - \lambda_b)\gamma s. \tag{45}$$

As in Brainard's analysis, only the fraction λ_b of the predicted effect of a change in the indicator is offset by optimal monetary policy. In contrast to this, (41) implies that all of the predicted effect will be offset by the outcome of monitoring. Thus optimal monetary policy without monitoring can in this case accomplish more than monitoring. With private monitoring the real money supply will overreact to information about aggregate disturbances. Specifically, it follows that the social gain to monitoring when $g = g'$ is

$$V^1(g') - V^2(g') = (k/2)\lambda_c\sigma_x^2\sigma_b^2[1/(\beta^2 + \sigma_b^2) - 1/\beta^2] < 0. \tag{46}$$

Optimal monetary policy can be characterized in this variant model as we did in the basic model with results that are similar but with a complication, because even when $g = g'$ and the social gain to monitoring is negative, the private gain in this case is actually positive. Specifically,

$$V^3(g) - V^2(g) = (k/2)\lambda_c\sigma_x^2(1 - \lambda_b g/g')^2,\text{[18]} \tag{47}$$

[18] See Appendix.

from which it follows that $V^3(g') - V^2(g') > 0$. Thus optimal monetary policy may require g to be somewhat larger than g' so that people will not be motivated to monitor.

For our purposes the important feature of optimal monetary policy is that the activist policy ($g = g'$) will be optimal whenever

$$(k/2)\lambda_c\sigma_x^2(1 - \lambda_b)^2 \leqq z_p + z_c < (k/2)\lambda_c\sigma_x^2 \tag{48}$$

and

$$z_a \leqq (k/2)\lambda_c\sigma_x^2\sigma_b^2[1/\beta^2 - 1/(\beta^2 + \sigma_b^2)]. \tag{49}$$

If (48) is satisfied then monitoring will occur under a neutral or disseminating policy but not under an activist policy. If (49) also holds then the activist policy is optimal because its cost (relative to $V^2[g']$) is z_a, whereas, from (46) and the fact that $V^1(g)$ is independent of g, the cost of each of the other policies will exceed the RHS of (49) by the sum of monitoring and administrative costs. As in the previous section, the monetary authority needs no cost advantage for these conditions to be satisfied if $\lambda_b < 1$. For example, they may be satisfied even if $z_a > z_c + z_p$, because the first term in (48) is just λ_b times the RHS of (49).

Thus, an activist monetary policy can be justified because the monetary authority, by avoiding the overreaction to the indicator that would occur with monitoring, is able to accomplish more than could be accomplished by private monitoring. The overreaction of private agents monitoring s occurs because each individual sees himself as being able to offset the effect of s by changing his relative price, the coefficient of which, n, is known with certainty. Thus, he will try to offset completely the predicted effects of a change in s. As Brainard has shown, this is the optimal policy when the effect of the price setter's action can indeed be predicted with certainty. But the effect of every agent trying to alter his relative price is a change in the absolute price level, the effect of which cannot be predicted with certainty. As Brainard has also shown, the optimal reaction when the effect of policy is uncertain is to attempt to offset less than 100 percent of the predicted effect of the disturbance. The crucial assumption in this argument is that the degree of uncertainty attached to the effects of relative-price changes is different from that attached to absolute-price changes. Private agents take the former into account but not the latter. In the particular model used above the former uncertainty is less than the latter, but this is not essential to the argument. In the reverse case private agents would *under*react to monitored aggregate disturbances, once again giving a rationale to an activist policy.

VI. Conclusion

In summary, this paper has argued that an activist monetary policy may be justified even when people's expectations are formed rationally if there are costs of gathering and processing information and if there is uncertainty concerning the exact values of parameters in the economic system. The justification that we have presented is twofold: first, that private agents may not be motivated to use information that can be used profitably by the monetary authority and, second, that even if they are motivated to use it they may not make as good use of it as could the monetary authority.

The formal argument has been presented with a simple model employing specific functional forms chosen for their tractability. Therefore, much work remains before the degree of generality of the results can be ascertained.

The first rationale that we provided for an activist policy should be distinguished from that of Fischer (1977) and Phelps-Taylor (1977). Their models are similar to ours in that prices are set in advance of trading rather than being determined by an independent auctioneer according to market-clearing conditions. However, their models allow activist monetary policies to improve welfare because they postulate a temporary rigidity in prices that prevents private agents from realizing all mutually advantageous and perceived gains from trade. This raises the question of whether the contractual arrangements implicit in such models would be competitively sustainable (see, e.g., Barro 1977). Our argument has gone one step further and shown conditions under which these contracts are in fact sustainable. We have not postulated any rigidity or imperfection that allows the monetary authority to react more quickly or with more flexibility to variation in the indicator than private agents can. Instead we have shown that private agents may *choose* not to act on as much information as the monetary authority possesses, or even as much as the monetary authority makes freely available to them, even when it would be Pareto improving for them to do so. In other words, this argument also provides an explanation for the fact that private contracts are typically written in nominal terms without being indexed to the money supply or other aggregate variables.

The paper has not addressed the question of whether or not the particular conditions derived above for the optimality of an activist policy actually exist in any real-world situation. The purpose has been mainly to argue that as a matter of principle the assumption of rational expectations does not imply the uselessness of an activist policy. It depends at least on the factors discussed above. Any bal-

anced evaluation of alternative monetary policies should also take into account at least two other important considerations. First, as Hayek (1945) has argued, the most important problem to be resolved is probably not what policy to pursue but whom to entrust with the task. An activist policy will work only if the authority can be entrusted or persuaded to pursue the public interest instead of his own and to pursue it competently. Second, as Hayek has also argued, the centralization of decision making is most easily rationalized when the decisions require the use of general knowledge rather than local, specific knowledge. In the model of the present paper every individual agent is essentially identical, so there is no clear distinction between these two kinds of knowledge. But more generally one would expect the aggregate disturbances to which monetary policy reacts to come under the heading of general knowledge, about which most individual agents can be expected to have little expertise. Thus a further argument in favor of an activist policy is that it allows society to centralize, and thus to bring to bear an efficient degree of expertise on, decisions concerning such general knowledge.

Appendix

This Appendix sketches a formal derivation of the private value of monitoring. In general terms, suppose that a decision maker has an information set I and his problem is to choose the value of decision variable x so as to

$$\max_{\{x\}} E(\alpha_0 + \alpha_1 x + \alpha_2 x^2 \mid I) \tag{A1}$$

where the α_i's are random parameters. Then the expected value of receiving the information contained in the new information set I' is the expected difference in the optimized value of (A1) resulting from replacing I by I'; that is,

$$E\{[E(\alpha_1 \mid I)]^2/4E(\alpha_2 \mid I) - [E(\alpha_1 \mid I')]^2/4E(\alpha_2 \mid I')\}. \tag{A2}$$

1. To derive (26) from (A2) note that in this case $\alpha_1 = (q_n^2 + n)(1 + nk)$, $\alpha_2 = -(n/2)(2 + nk)$, $I = \emptyset$, and $I' = \{s\}$.

2. To derive (28), let x be the *average* relative price, $PE[P_n^1(\overline{M}, g, s)^{-1}]$, and note that in this case $\alpha_1 = e_n(q_n^1 + n)(1 + nk)$, $\alpha_2 = -e_n^2(n/2)(2 + nk)$, $I = \emptyset$, and $I' = \{s\}$.

3. To derive (48) proceed exactly as in deriving (26) and take the limit as $n \rightarrow \infty$.

References

Barro, Robert J. "Rational Expectations and the Role of Monetary Policy." *J. Monetary Econ.* 2 (January 1976): 1–32.

———. "Long-Term Contracting, Sticky Prices, and Monetary Policy." *J. Monetary Econ.* 3 (July 1977): 305–16.

Brainard, William C. "Uncertainty and the Effectiveness of Policy." *A.E.R. Papers and Proc.* 57 (May 1967): 411–25.

Fane, G. "Stabilization Policy in Models with Rational Expectations and Uncertainty." Mimeographed. Australian National University, 1977.

Feige, Edgar L., and Pearce, Douglas K. "Economically Rational Expectations: Are Innovations in the Rate of Inflation Independent of Innovations in Measures of Monetary and Fiscal Policy?" *J.P.E.* 84, no. 3 (June 1976): 499–522.

Fischer, Stanley. "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule." *J.P.E.* 85, no. 1 (February 1977): 191–206.

Frenkel, Jacob A. "Inflation and the Formation of Expectations." *J. Monetary Econ.* 1 (October 1975): 403–21.

Friedman, Benjamin M. "Optimal Expectations and the Extreme Information Assumptions of 'Rational Expectations' Macromodels." *J. Monetary Econ.* 5 (January 1979): 23–41.

Friedman, Milton. "The Effects of a Full-Employment Policy on Economic Stability: A Formal Analysis." In *Essays in Positive Economics.* Chicago: Univ. Chicago Press, 1953.

———. "The Role of Monetary Policy." *A.E.R.* 58 (March 1968): 1–17.

Grossman, Sanford J., and Stiglitz, Joseph E. "Information and Competitive Price Systems." *A.E.R. Papers and Proc.* 66 (May 1976): 246–53.

Hayek, Friedrich A. von. "The Use of Knowledge in Society." *A.E.R.* 35 (September 1945): 519–30.

Hirshleifer, Jack. "The Private and Social Value of Information and the Reward to Inventive Activity." *A.E.R.* 61 (September 1971): 561–74.

Laidler, David E. W. "Money and Money Income: An Essay on the 'Transmission Mechanism.' " *J. Monetary Econ.* 4 (April 1978): 151–91.

Lucas, Robert E., Jr. "Expectations and the Neutrality of Money." *J. Econ. Theory* 4 (April 1972): 103–24.

———. "An Equilibrium Model of the Business Cycle." *J.P.E.* 83, no. 6 (December 1975): 1113–44.

Phelps, Edmund S., and Taylor, John B. "Stabilizing Powers of Monetary Policy under Rational Expectations." *J.P.E.* 85, no. 1 (February 1977): 163–90.

Poole, William. "Rational Expectations in the Macro Model." *Brookings Papers Econ. Activity*, no. 2 (1976), pp. 463–505.

Rutledge, John. *A Monetarist Model of Inflationary Expectations.* Toronto: Lexington, 1974.

Sargent, Thomas J. "A Classical Macroeconometric Model for the United States." *J.P.E.* 84, no. 2 (April 1976): 207–37.

Sargent, Thomas J., and Wallace, Neil. " 'Rational' Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule." *J.P.E.* 83, no. 2 (April 1975): 241–54.

Shiller, Robert J. "Rational Expectations and the Dynamic Structure of Macroeconomic Models: A Critical Review." *J. Monetary Econ.* 4 (January 1978): 1–44.

Taylor, John B. "Monetary Policy during a Transition to Rational Expectations." *J.P.E.* 83, no. 5 (October 1975): 1009–21.

Equalizing Discrimination and Cartel Pricing in Transport Rate Regulation

Kenneth D. Boyer

Michigan State University

There are two possible outcomes of transport regulation: (1) maintaining a carrier cartel and (2) imposing equalizing discrimination against advantaged and in favor of disadvantaged shippers. Both functions have required a complex rate structure to enforce the respective forms of price discrimination. Using a sample of freight bills from motor carriers and railroads, this paper demonstrates that the principal result of motor carrier regulation has been to maintain a cartel of truckers, while railroad regulation has thwarted the wishes of the railroad cartel by imposing equalizing discrimination on weak and strong shippers. Motor carrier rates respond in an economically rational manner to costs and shipper bargaining power; rail rates are either unresponsive or perversely responsive to the same factors. Deregulation should have divergent effects in the two industries.

Deregulation is a policy that has been recommended in recent years for some but not all regulated industries (see, e.g., Council of Economic Advisers 1979). Candidates for deregulation are most often in the transport sector and never in the traditional public utilities such as electricity and natural gas distribution. Why is transport regulation consistently singled out for attention? Also curious are the extraordinary attempts of motor carriers to retain the regulatory fetters, while the railroads make somewhat half-hearted attempts to remove theirs (see, e.g., Williams 1978). What is it about the type of regulation that makes

This paper was improved by advice and comments from W. G. Shepherd, A. S. Lang, H. Trebing, G. Wilson, P. Schmidt, R. Rasche, D. Suits, S. Tainter, and an anonymous referee.

[*Journal of Political Economy*, 1981, vol. 89, no. 2]

one transport industry wish to retain—at all costs—what another transport industry is willing to shed?

As this paper explains, the answer to both questions is found in the institution of transport rate making. There are two key differences between traditional forms of price regulation and transport rate controls. First, in most instances, regulation involves category pricing, while transport rates are tailored to individual buyers. Second, transport rates are typically the result of collective rate making, while most other regulated prices are not. Together, these aspects allow the regulatory authority to bestow advantages and disadvantages on individual communities and industries; they also allow the industry to stabilize a sellers' cartel.

This paper demonstrates that there would be two distinct results from deregulating the motor carrier and railroad industries since there are divergent outcomes of regulation in the two modes. Current motor carrier rates show classic signs of cartel rate making; railroad rates reveal Interstate Commerce Commission (ICC) attempts to equalize advantages and disadvantages of shippers and localities. Deregulating the trucking industry would inhibit cartel rate making; deregulating the railroad industry would lead to a more economically rational pattern of prices, with some rates rising and some falling.

To emphasize the unusual nature of transport pricing, Section I shows what category pricing would be like in transport. Section II shows how individualized rates and collective rate making lead to unpredictable outcomes in the regulatory process. Unlike category pricing, transport rate making is ideally suited for either cartel pricing or enforcing equalizing discrimination. Section III investigates the current motor carrier and railroad rate structures and finds that different principles are apparent in the pricing of the two modes. The paper concludes with implications of the findings for deregulation arguments.

I. Transport Prices Are Individual Rather than Category Rates

Compared with motor carriers and railroads, the rate structure for most regulated industries is the model of simplicity: The customer's class of service is determined, and the block pricing for the class is applied. Such category pricing has two important implications. First, a customer's charges can be raised or lowered only by altering the prices assessed everyone in that category. Customers cannot demand special price concessions and are protected from unusually high rates by the presence of other members in the category. Category pricing prevents bargaining over individual charges between suppliers and customers.

The second implication is for regulators. Because the pricing is

relatively simple, the regulatory authorities can make informed decisions about each price. The costs of providing service to a category of customers can, at least in principle, be calculated with some accuracy; again in principle, the regulators can check the supplier's cost calculations with a limited staff.

The costs of providing transport service are quite variable (Harris 1977; Spady 1979). Thus category prices for railroads and motor carriers would be more complicated than the price structures of most regulated industries. Such a scheme would, nonetheless, be feasible. For example, railroad car rentals (which could change daily according to availability) could be separated from haulage charges. Haulage charges could be divided into categories depending on the number of cars in the train, the regularity of movement, and so forth. Track rentals could also be charged and graduated according to expected maintenance charges on a link traversed. Congestion tolls could be assessed if a shipment used a link with excessive traffic. Separate charges could be assessed for using switching locomotives and renting yard space.

Similarly, motor carriers could be compelled to charge separate prices for separate services: one price for pickup, another for terminal sorting, another for use of line-haul service on a certain link, and so on. Total transport charges would then be determined by summing the costs of the service categories used.

As long as a transporter could charge for only a limited number of service categories, this scheme would mean that a customer's charges could be altered only by changing the rates of all those in a category, and that, by limiting the number of prices reviewed by the ICC, a competent staff could determine the relation between the price and the cost of a service. As demonstrated below, if transport regulation had the nature of category pricing, there would be scant reason to expect intense political pressure either to preserve or to remove the regulatory apparatus.

II. The Implications of Pricing Services Individually

Most transport sales are not based on category prices but use a scheme of exception or commodity rates. Each price is tailored to the individual buyer (Locklin 1972). As demonstrated below, this has inevitable consequences for the relations among carrier, customer, and regulator.

A system of class rates for motor carriers and for railroads does exist. Under this scheme, each commodity is given a class rating and each origin-destination pair is given a rate base number. The transport price is the product of these two numbers. The class-rate tariff is

thus a two-category scheme. It makes pricing simple and thus capable of intelligent oversight by the ICC; it also means rates cannot be changed for one shipper without simultaneously changing them for all others in the same class.

Because the class-rate scheme is only a two-category system, it cannot adequately reflect carrier costs which vary in more than two dimensions. In particular, neither economies of traffic density nor the likelihood of receiving a backhaul is accounted for in the class-rate system.

These unusual cost circumstances can be accommodated through the commodity-rate and exception-rate schemes which allow the quotation of a special price for shipping a good defined at greater than seven-digit precision between any two stations. That rates can be quoted so precisely reveals that they apply in practice to individual shippers. Commodity and exception rates are the opposite of category rates—they are individual rates.

Commodity and exception schemes are now almost the universal pricing method for rail shipments, and they are the predominant form of truckload rates. The class-rate system is reserved for occasional shippers and for much less-than-truckload traffic.

To understand the nature of transport rate regulation is to understand the interaction between the ICC and those parties that propose a change in a commodity or exception rate. More than 85 percent of motor carrier rate changes are reductions proposed by carriers or, more frequently, the rate bureaus to which they belong.[1] The universal rationale is to retain or gain traffic that would otherwise be moved by another mode.

Under individual pricing, shippers may gain advantages over competitors if their own rates are lowered but other shippers' are not. To bring this about, a shipper must convince the carrier that either (1) existing traffic will be diverted without the lower rate or (2) new, formerly uneconomical markets will be opened up if the rate is lowered. A shipper has two hurdles: First, the shippers must convince the carriers who would offer the rate that the threat of traffic gain or loss is genuine;[2] second, should the rate be protested, the ICC must rule

[1] This figure was quoted by M. A. Godecker of the Central States Motor Freight Bureau. My understanding of regulatory procedure in this paper benefited by discussions with Godecker and J. R. Ahlstrom of the Western Truck Line Committee. Substantial alterations in the procedures for regulating railroads and motor carriers were passed by Congress in the time subsequent to the writing of the paper. The forms of regulation discussed in this paper were those practiced in through the spring of 1980.

[2] Motor carriers claim that collective rate making is necessary to protect them from shippers' opportunistic misrepresentation of the likelihood of traffic diversion.

on it. To go into effect the rate must be declared just, reasonable, and nondiscriminatory.

Precisely what criteria the ICC uses to judge rates have never been defined. The commission does demand cost evidence in the form of standard cost statements. It recognizes, however, that the costs summarized in the statements are average costs, and thus additional evidence of haulage costs is used; the standard piece of evidence is the rate of analogous hauls.

The basic ICC regulatory procedure assures a balance of rates on analogous shipments. A just, reasonable, and nondiscriminatory rate is one which is not out of line with rates on analogous hauls. What is in practice regulated, then, is the rate relationship among analogous or competing hauls.

The result of this regulatory procedure has been a transport rate structure of fantastic complexity: Hundreds of billions of individual prices have been established and approved. In contrast, under the typical category scheme used in most regulated industries, there are rarely more than 100 separate rate categories. The complexity of the transport rate structure precludes staff investigations of price-cost relationships for individual rates. Indeed, the very complexity of the rate structure discourages cost finding. With traffic equal to the sum of large numbers of unique movements, virtually all costs will be found to be joint or common and thus logically or practically untraceable.

This inability to cost individual movements has distinct advantages both to carriers and to the ICC since it increases the discretion of each party to pursue separate goals in the regulatory process. The lack of reliable cost figures frees the rate bureaus to engage in demand elasticity–based pricing which is implicit in the policy of granting rate reductions only to those shippers who can credibly threaten traffic diversion. Without cost data, discrimination is extraordinarily difficult to prove. The complaint of a firm that lower rates are charged to a rival can be answered by the apology that the rate concession was necessary to retain traffic. It is, of course, crucial to the proper working of the scheme that price-induced traffic diversions to firms with identical cost structures be inhibited, otherwise the process of collective rate making will evolve toward cost-based prices.

The difficulty of relating prices to costs gives wider latitude to the regulatory authorities to pursue what Hilton (1972) has called "the basic behavior of regulatory commissions." Hilton identified the goal of regulation as dissipation of monopoly profits through cross-subsidization; a preferable description in the case of transport would be requiring discrimination which equalizes the advantages and disadvantages of different purchasers of transport service. The phrase

"equalizing discrimination" is preferred to the more familiar title of "cross-subsidization" since the latter implies that the carrier uses profits from one service to finance another. But if carrier operations generally run a deficit, no service is paying for any other. The term "equalizing discrimination" emphasizes that the ICC is interested in equalizing the conditions of advantaged and disadvantaged shippers rather than in having one party pay the bills of another.

Equalizing discrimination is the natural result of attempting to prevent "unjust" or "extractive" discrimination (i.e., pricing to extract rents from shippers and localities) without using a cost comparison. For example, the volumes generated by a large shipper typically will produce both lower carriage costs and many eager carriers, while a small shipper will have fewer alternatives and higher carriage costs. To prevent exploitation of the small shipper's lower demand elasticity, in the absence of a reliable cost comparison, the regulatory authorities have no option other than to require equal prices for both shippers. The burden of this equalizing discrimination naturally falls on the advantaged shipper.

Explicit examples of equalizing discrimination can be found in policies that require identical prices for shipment of goods from the Midwest to any East Coast port regardless of distance, require low rates on California citrus fruit and high rates on Florida produce so that both can compete in Northeast markets, and discourage the quotation of unit train rates even when there are obvious cost advantages (MacAvoy and Sloss 1967).

The pressures on the ICC to practice such equalizing discrimination are most intense in the case of relatively low value, homogeneous commodities such as agricultural products, coal, steel, and paper. Demand for these goods is most sensitive to delivered prices, and their profitability is most strongly affected by transport prices. These are also the goods carried predominantly by rail. In contrast, the ICC gets less intense pressure to use transport prices to equalize the conditions of manufacturers of finished goods—the most important customers of motor carriers. This is because demand is less sensitive to transport prices and because transport rates are a smaller part of delivered cost.

Not only is there less pressure on the ICC to produce equalizing discrimination for those commodities hauled primarily by motor carriers, but the commission has less leeway, since carriers can use their own trucks if common-carrier rates get too high. One would expect, therefore, that the ICC would give more care to regulating rail than motor carrier rates: Their range of actions and the participants' interests are greater than in truck rates.

In the next section the rate structures of the two modes are investi-

gated to test the hypothesis that the regulatory influence of the ICC is more evident in railroad than in motor carrier pricing.

III. Empirical Evidence of Different Regulatory Regimes for Railroads and Motor Carriers

To demonstrate the dual outcome of ICC regulation, individual rates were regressed on factors which proxy demand elasticities and costs of providing service. A cartel unthwarted by the ICC will have rates that vary directly with the costs of service and inversely with demand elasticities.

It is a good deal more difficult to specify what would determine rates under equalizing discrimination. However, two aspects are likely: First, cost savings of a large shipping volume will not be passed along to customers in the form of lower charges; and second, having fewer shipping options is likely to lower rather than to raise rates since a low demand elasticity indicates a vulnerable shipper.

The complexity of rail and motor carrier rate structures derives from the discriminatory aims of both the rate bureaus and the regulators. But this same complexity precludes investigating rate structures in any manner other than explaining the patterns of actual charges for purchased services.[3] A sample of rail freight bills has long been available from the ICC, but the similar data file for motor carriers is the proprietary information of motor carrier rate bureaus.[4] This study was made possible by obtaining the latter file for the year 1972 and comparing it with the similar rail file.

The six analyzed commodities (the same for both rail and motor carriers) are perhaps not typical of either motor carrier or rail traffic as a whole; the two modes do not have identical traffic mixes.[5] Rather, the commodities were chosen to be representative of the goods contested by both railroads and motor carriers.

[3] The use of a sample of actual freight bills to model the rate structure rather than the alternative procedure of using rates printed in formal tariff statements is preferred for several reasons: (1) There are always numbers of rates available to any shipper, and it is not clear which would be the governing rate; (2) charges are sometimes different from those printed in tariffs; and (3) a large part of the rate structure is nominal in the sense that no freight is ever moved under these rates.

[4] The rail data are the ICC "One Percent Waybill Sample of Carload Terminations" (U.S. Interstate Commerce Commission 1972). The motor carrier data are from the "All-Bureau Continuing Traffic Study for 1972" (Rocky Mountain Motor Freight Bureau 1972).

[5] The six commodities are: prepared feed NEC, assigned STCC numbers 2042129 and 2042190 and NMFC number 72910; wrapping paper, STCC 2084120/21, NMFC 111510, 2621490, 151800; pulpboard or fiberboard, STCC 2631115/16120, NMFC 151310; corrugated fiberboard or pulpboard boxes, STCC 2651141, NMFC 29275; tires, STCC 3011110, NMFC 157230. For a complete description of commodity selection procedures and data filters, see Boyer (1978).

The charges for shipping a hundredweight of each of the six commodities between various end points were studied. One of the strongest influences was found to be the size of the shipment. In the Appendix, I describe the procedure used to adjust the reported charge per hundredweight in each freight bill for variations caused by differences in loading weight. The dependent variable whose variation is explained by the regressions in this section is a least-squares forecast of what the charge per hundredweight on a freight bill would have been had the shipment been made at a standard size of 340 cwt for trucks and 680 cwt for trains.[6] Only boxcar movements of rail commodities were considered.

The independent variables which could be generated from the freight bill samples were the length of the haul (*MI*); the number of shipments of this commodity, on this haul, in the sample (*SHP*); the number of alternative destinations open to a shipper of this commodity from this origin (*DST*); the number of origins for this commodity from which a receiver at this destination can choose (*ORN*); and existence of active competition from the other regulated mode of transport (*INT*). It can be assumed that costs per hundredweight increase with distance and decrease with shipping volume. Similarly, large values for *DST* and *ORN* as well as a positive value for *INT* should indicate a shipper with a higher demand elasticity. In addition, the existence of private trucking suggests that shipping volume will be positively related to demand elasticities.

The functional form for these regressions was:

$$\begin{aligned}\ln(\mathrm{RATE}_i) = {} & b_1(\ln MI_i) + b_2(\ln MI_i)^2 + b_3(\ln MI_i)^3 \\ & + b_4M(\ln MI_i - \ln 100)^3 + b_5N(\ln MI_i - \ln 500)^3 \\ & + C_1 \ln(SHP_i) + C_2 \ln(DST_i) + C_3 \ln(ORN_i) + C_4(INT) \\ & + \sum_{j=1}^{n} d_j D_{ji} + e_i,\end{aligned}$$

where

RATE = cents per hundredweight shipping charges (see Appendix);
MI = length of haul;
M = 1 if $MI_i > 100$, 0 otherwise;
N = 1 if $M_i > 1{,}000$, 0 otherwise;
SHP = number of freight bills in sample corresponding to this haul and commodity;

[6] These correspond to truckloads of 17 tons and carloads of 34 tons which were taken to be representative in Meyer et al. (1959).

DST = number of different destinations for freight bills in sample with this origin and commodity;

ORN = number of different origins for freight bills in sample having this destination and commodity;

INT = 1 if there is a freight bill in both samples with this commodity, destination, and origin and 0 if there is no intermodal competition; and

D_{ji} = 1 if observation is for commodity j and 0 if observation is not for commodity j.

The functional form contains a cubic spline of distances with knots placed at 200 and 1,000 miles.[7] I used this piecewise cubic approximation of the distance taper to avoid prior constraints on the shape of the distance relationship.

The regressions in table 1 show the results for the motor carrier rate structure. Table 1 also contains two rail regressions, one for the complete set of six commodities and one for a restricted set of the four goods for which haulage weight/rate relations could be estimated (see Appendix).

As reported in table 1, the results strongly support the conjecture that motor carriers use a cartel rate structure unhampered by government interference. Large shippers do get lower rates (the elasticity of rates with respect to the number of shipments is −.039) as predicted by the cartel model; the cost of serving large shippers is lower and the likelihood of switching to private trucking is higher. The coefficient on the number of destinations used by the seller (DST) and the number of origins used by the receiver of a shipment (ORN) similarly confirm that the greater the number of competing hauls, the lower the rates. Finally, the coefficient on INT suggests that motor carrier rates are approximately 6 percent lower where there is actual competition from railroads than where there is not. All of these findings are consistent with cartel rate making in the trucking industry.

The contrast to the rail results is striking. The response of rail rates to larger shipping volume (SHP) is insignificant. A large number of destinations (DST) lowers a shipper's rates, but a larger number of origins (ORN) does not similarly lower rates. And, finally, the presence of motor carrier competition (INT) is associated with higher, not lower, rail rates. With the exception of the negative sign on DST, the remaining results are those predicted by a policy of equalizing discrimination where rates are either unresponsive or perversely responsive to cost conditions and demand elasticities. The negative

[7] On estimating spline functions with ordinary least squares, see Suits, Mason, and Chan (1978).

TABLE 1

DETERMINANTS OF REGULATED FREIGHT RATES, CUBIC SPLINE REGRESSIONS

	Motor Carrier, Eq. (1)		Rail Eq. (2)		Rail Eq. (3)	
ln *MI*	3.85	(3.71)	−17.35	(4.53)	−20.64	(4.87)
$(\ln MI)^2$	−.77	(.73)	3.16	(.89)	4.11	(.96)
$(\ln MI)^3$	.05	(.04)	−.22	(.058)	−.26	(.06)
$(\ln MI - \ln 100)^3$	−.04	(.06)	.22	(.074)	.27	(.079)
$(\ln MI - \ln 500)^3$	−1.92	(.10)	.23	(.10)	.27	(.11)
SHP	−.037	(.014)	−.011	(.012)	−.008	(.012)
DST	−.0062	(.0042)	−.014	(.006)	−.014	(.006)
ORN	−.012	(.1153)	.0018	(.0070)	−.00031	(.0071)
INT	−.062	(.021)	.060	(.020)	.064	(.021)
Dummy variable:						
3011110	.12	(.0079)	.46	(.015)	.48	(.015)
2651141	.26	(.016)	.51	(.028)	...	...
2631115	−.29	(.013)	.067	(.0090)	.09	(.008)
2621490	−.12	(.013)	.13	(.017)	.15	(.017)
2084120	−.12	(.037)	−.11	(.066)	...	...
2042190	−.24	(.016)	−.61	(.014)	−.58	(.014)
R^2	.54		.82		.83	
N	6,472		4,062		3,836	
F	658.78		904.06		855.56	

SOURCES.—U.S. Interstate Commerce Commission 1972; Rocky Mountain Motor Freight Bureau 1972.
NOTE.—Dependent variable = fitted cents/hundredweight for end-point pairs. Numbers in parentheses are SEs.

coefficient on *DST* indicates that there is some rational response to demand elasticities as well—an indication of some cartel rate making within a structure whose primary characterization is enforcing equalizing discrimination.

The regressions in table 1 give other evidence consistent with the hypothesis that railroads are thwarted by regulatory authorities while motor carriers set rates just as a successful trucker's cartel would wish. The first evidence comes from the spline variables on mileage. The mileage tapers, expressed on the vertical axis as a percentage of charges for a 500-mile haul, are graphed in figure 1.[8] The two modes appear to be very similar; the rail taper has unusually low short-haul rates and relatively high charges for the longest hauls. This is precisely the opposite of what would be generated by cost-based rates and contradicts the basis for many studies of optimal intermodal traffic allocation (Friedlaender 1969; Levin 1978). The slopes and positions of the two curves indicate that the observed propensity for long hauls

[8] On the safe assumption that rail rates will be lower than road rates at the standard 500-mile haul, the results suggest that rail and truck rates either diverge as mileage increases or that the rail curve cuts the motor carrier rate curve from below.

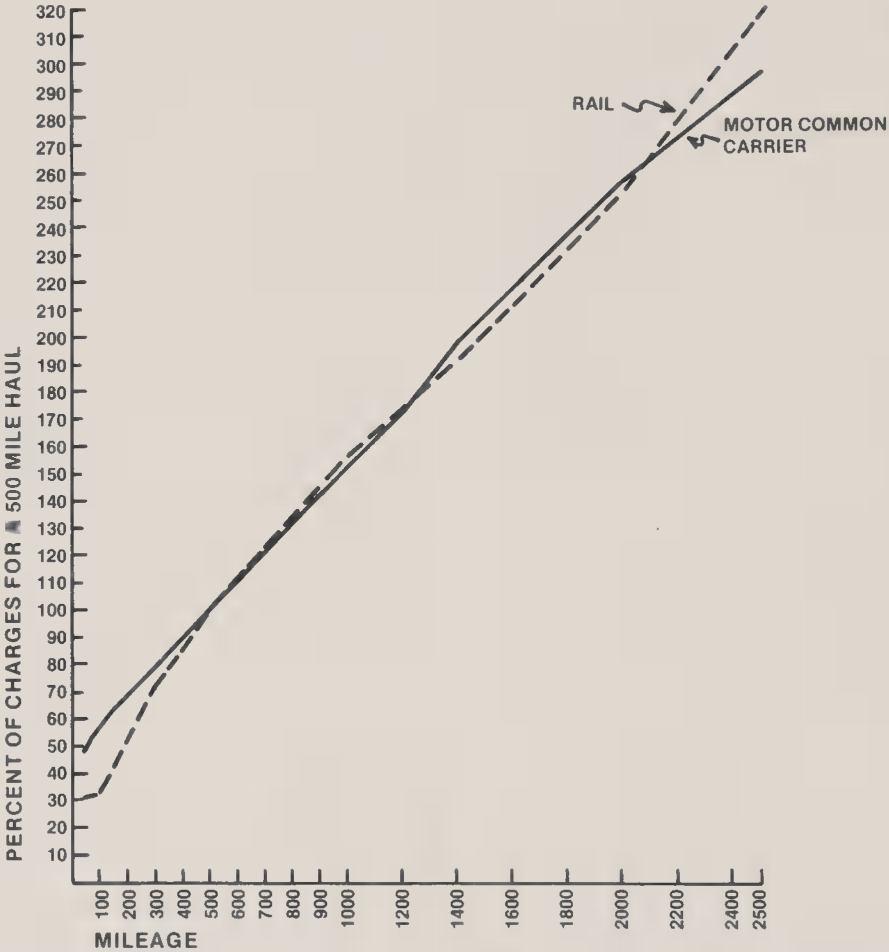


Fig. 1.—Mileage tapers for rail and motor carriers. Source: spline coefficients estimated in table 1.

to be shipped by rail and short hauls by truck is not the result of shippers responding to cost-based rates but, rather, must be due to a convergence of service qualities.

The results are consistent with an ICC policy of "fair sharing traffic" among modes. Rail rates are held high for long hauls to allow motor carrier participation and are allowed to fall for short hauls where motor carriers do not have to fear serious traffic diversions.[9]

A final piece of evidence for the differing regulation of motor carrier and railroad rates comes from the commodity-specific dummy variables in table 1. Converting logarithmic coefficients to percentages yields the data in table 2. The numbers are average rates for each of the six commodities. The variation in rail rates explainable by com-

[9] On fair sharing of traffic, see the "Ingot Molds Case" (American Commercial Lines, Inc., et al. v. Louisville and Nashville Railroad Co. et al. 1968). Note that the evidence is also consistent with one or the other, but not both, of the modes of pricing under cartel rate making.

TABLE 2

VALUE OF SERVICE PRICING

Commodity	Percentage of Charges on Average Commodity: Motor Carrier from Eq. (1)	Rail: From Eq. (2)	Rail: From Eq. (3)
3011110	113.22	159.44	163.02
2651141	130.31	167.35	...
2631115	74.41	106.97	109.46
2621490	87.96	114.37	116.72
2084120	88.58	88.95	...
2042190	78.01	53.80	55.49
SD	21.82	42.90	44.05

SOURCE.—Table 1, antilog of commodity dummy variables.

modity value is more than twice as high as that for motor carrier rates: The column standard deviations are 42.9 percent for rail as opposed to 21.8 percent for truck. This suggests that rail rates are more closely related to commodity type than are truck charges.[10] This finding is consistent with a railroad industry which maintains an obsolete system of value-of-service pricing and a trucking industry which has adapted to an environment of intermodal competition.

IV. Conclusion

The evidence in this paper shows that both motor carrier and railroad rates are different from those that would be observed without regulation. The pattern of rates suggests, however, that regulation of the two modes has had different results. Perhaps the strongest support for the hypothesis that railroads and motor carriers are subject to different types of regulation comes from the public posture of the two modes toward deregulation. Within the last years, the Association of American Railroads has endorsed a "deregulation" proposal. However, the proposal does not suggest that collective rate making be abandoned. In contrast, the American Trucking Associations is vehement in its opposition to any changes in transport regulations. This is consistent with the findings of this paper: (1) Motor carrier regulation ratifies a carrier cartel, while (2) rail regulation has classic equalizing discrimination as its primary characteristic.

The removal of government oversight over rates and entry and the simultaneous outlawing of collective rate making should have diver-

[10] The simple correlation between cols. a and b is .71.

gent outcomes in the railroad and motor carrier industries. A deregulated trucking industry would—at least in truckload service—develop a rate structure more closely related to costs of service. A general reduction in trucking prices would be expected.

Deregulating rail rates would have far more complex results. Given the small number of rail carriers in any market, a strictly cost-based rate structure could not be predicted; nor, if rail marginal costs are below average costs, would this result be desirable. Instead, my analysis suggests that a more rational rail rate structure would result, one based both on costs and on the ability to divert traffic (the latter related, presumably, to trucking costs). Deregulating rail rates should not lead to a general increase or decrease in rail rates but to a reshuffling of prices.

This expectation is confirmed by the recent removal of ICC oversight over fruit and vegetable tariffs. The diversion of traffic to unregulated truckers over the last 2 decades effectively defeated the attempt at equalizing discrimination, and in 1979 rail rates on this traffic were deregulated. The result was a simultaneous increase of some rates and a sizable diversion of traffic to railroads in response to lower rates for other hauls (Association of American Railroads 1979; American Trucking Associations 1979). There is every reason to expect that deregulation of rail rates for other commodities would have a similar effect.

The disparate effects of regulation in the two modes are the direct result of the institution of pricing services individually rather than by categories. Individual pricing, which has allowed the ICC to pursue a policy of equalizing discrimination, has also required the use of rate bureaus and collective rate making. Had the transport industries used category rate making, regulation to enforce equalizing discrimination would have been hampered and collective rate making would have been unnecessary, and, therefore, it is doubtful that the fervent calls for and against deregulation would be heard today.

Appendix

The regressions in the text explain variations in fitted or standard rates for each commodity-origin-destination combination. These rates are adjusted for differences caused by variable loading weight among different commodities.

While most discussions of transport pricing proceed on the assumption that there is a single charge per unit of service, in fact this is not the case. There are separate charges per hundredweight for shipments of a commodity between two end points for (*a*) minimum charge, (*b*) less-than-truckload (LTL), (*c*) truckload (TL) or carload, and (*d*) very heavy shipments. Due to the different charging regimes there is an inverse relationship, "a weight taper," between charges per hundredweight and loading weight. To develop a

structure of single rates for commodities between end points, the first step was to estimate a weight taper for each commodity.

The weight tapers were estimated on the basic assumption that for every origin-destination pair, the shipping charges per hundredweight for a shipment of one size would be proportionate across different levels to the charges for a shipment of a standard vehicle load of the commodity. The standard procedure of spline estimation was used to estimate the shape of the taper in the absence of prior knowledge of its functional form. Knots were arbitrarily placed at 100 and 300 cwt for motor carrier estimation and at 500 and 1,000 cwt for railroads.

The functional form used to measure the rate taper with respect to loading weight for each commodity was:

$$\ln \text{cents}_i = b_1(\ln wt_i) + b_2(\ln wt_i)^2 + b_3(\ln wt_i)^3 + b_4F(\ln wt_i - \ln f)^3 + b_5G(\ln wt_i - \ln g)^3 + \sum_{j=1}^{n} c_jD_{ji} + e_i,$$

where

- cents_i = cents per hundredweight for shipment;
- wt_i = weight of shipment in hundredweight;
- D_{ji} = haul dummies: $D_{ji} = 1$ if the shipment is on end-point pair j, $D_{ji} = 0$ if the shipment is not on end-point pair j;
- F = 1 if $wt_i > 100$ for motor carrier or $wt_i > 500$ for railroads, 0 otherwise;
- f = 100 for motor carriers, 500 for rail;
- G = 1 if $wt_i > 300$ for motor carrier or $wt_i > 1{,}000$ for railroads, 0 otherwise;
- g = 300 for motor carriers, 1,000 for rail; and
- e_i = error term.

The values of the spline variables are listed in tables A1 and A2 for rail and motor carriers, respectively. The number of dummy variables is listed in each table as the number of hauls. For two commodities, there was insufficient variation in the independent variable to estimate the relationships for railroads.

In general, the tapers have at least one steeply downward-sloping segment which for motor carriers can be interpreted as the transition from LTL rates to TL rates. As should be expected from commodities that differ in loading characteristics, this steep downward-sloping section occurs at different weights for different commodities.

Some commodities have flat motor carrier tapers beyond the first downward-sloping segment, while others continue to fall; still others have a slightly upward-sloping segment at the very end. Since spline functions are piecewise cubic, it is not clear how much faith should be placed in the aberrant behavior of the tapers of some commodities for very large and very small shipment sizes.

The downward-sloping segment of the weight taper for rail-shipped commodities is found in most cases at much larger shipment sizes than for motor carriers. For most rail commodities, the downward slope is not as steep as for trucks. As must be expected for commodities with different loading characteristics, there are exceptions to this rule. Without exception, however, the downward slopes for rail commodities are at a much higher range than for truck commodities.

TABLE A1

MOTOR CARRIER WEIGHT TAPER REGRESSIONS

INDEPENDENT VARIABLE	COMMODITY					
	2042190 Feed	2084120 Wine	2621490 Wrapping Paper	2631115 Pulpboard or Fiberboard	2651141 Corrugated Boxes	3011110 Tires
ln WT	−.61	−1.13	17.64	−8.92	−60.98	−14.61
$(\ln WT)^2$	.31	.57	−4.31	2.14	15.33	3.76
$(\ln WT)^3$	−.04	−.08	.34	−.18	−1.26	−.32
$(\ln WT - \ln f)^3$	.20	.55	−.60	.41	2.63	.77
$(\ln WT - \ln g)^3$	.38	−1.11	1.13	−.17	−4.33	−1.38
R^2	.996	.998	.989	.981	.931	.909
r^2	.781	.946	.605	.455	.56	.331
Observations	731	128	1,101	1,165	764	4,409
Hauls	644	108	923	958	632	3,308
F	57.34	49.43	52.79	33.64	4.66	108.42

SOURCE.—Rocky Mountain Motor Freight Bureau 1972.

NOTE.—Dependent variable = ln (cents per hundredweight).

TABLE A2

RAIL WEIGHT TAPER REGRESSIONS

Independent Variable	2042190 Feed	2621490 Wrapping Paper	2631115 Pulpboard or Fiberboard	3011110 Tires
b_1	172.48	−103.78	−128.61	−7.47
b_2	−30.02	17.61	21.13	1.10
b_3	1.72	−.99	−1.16	−.05
b_4	−3.81	1.85	1.62	−3.71
b_5	5.11	1.69	−.82	251.37
R^2	.963	.998	.980	.979
r^2	.08	.90	.378	.296
Observations	1,187	562	3,159	999
Hauls	890	432	1,941	577
F	5.38	231	147.63	35.12

SOURCE.—Interstate Commerce Commission 1972.

NOTE.—Dependent variable = ln (cents per hundredweight). Equations for wine and boxes could not be estimated due to insufficient variation in the independent variables.

Two coefficients of multiple determination are reported for each equation. As usual, R^2 is interpreted as the percentage of the total variation in the dependent variable which can be explained by the right-hand-side variable. This is misleading, however, since the extensive use of dummy variables accounts for much of the apparent explanatory power of the regressions. The second coefficient of multiple determination, r^2, is defined as:

$$r^2 = \frac{\sum_{i=1}^{n_j}\sum_{j=1}^{m}(Y_{ji} - \bar{Y}_j)^2}{\sum_{i=1}^{n_j}\sum_{j=1}^{m}(Y_{ji} - \bar{Y}_j)^2 - \sum_{i=1}^{n_j}\sum_{j=1}^{m}(\hat{Y}_{ji} - \hat{\bar{Y}}_j)^2},$$

where Y_{ji} is the ith observation on the dependent variable in group j, $\bar{Y}_j$ is the mean of the dependent variable for observations in group j, and $\hat{Y}_{ji}$ is the fitted value of the ith observation on the dependent variable in group J. Numbers in the row marked r^2 are thus a more accurate indication of the explanatory power of the five spline variables. The reported F-statistic is computed using r^2.

The dummy variables in each equation were computed using the analysis of covariance. They were evaluated, however, not at 0 cwt but at 340 cwt for trucks and 680 cwt for trains. The dummy variables thus are interpreted as a standard rate for each haul and commodity: a least-squares forecast of the rate had all shipments been made at the standard size. The standard rate thus purges all influence of variable loading weights within a commodity. For the two rail commodities whose weight tapers could not be estimated, the standard rate was computed as the average charge for the haul, regardless of loading weight.

References

American Commercial Lines, Inc., et al. v. Louisville and Nashville Railroad Co. et al. "Ingot Molds Case." 392 U.S. 571 (1968).

American Trucking Associations. "Decontrolled Rail Movements of Fresh Fruits and Vegetables." *Res. Rev.*, no. 215 (October 15, 1979), pp. 2–3.

Association of American Railroads. "Railroads Reverse Decline in Fruit, Vegetable Traffic." *Information Letter*, no. 2280 (September 26, 1979), pp. 1–4.

Boyer, Kenneth D. "On the Structure of Regulated Freight Rates." Econometrics Workshop Paper no. 7801, Michigan State Univ., August 1978.

Council of Economic Advisers. *Economic Report of the President.* Washington: Government Printing Office, January 1979.

Friedlaender, Ann E. *The Dilemma of Freight Transport Regulation.* Washington: Brookings Inst., 1969.

Harris, Robert G. "Economics of Traffic Density in the Rail Freight Industry." *Bell J. Econ.* 8 (Autumn 1977): 556–64.

Hilton, George W. "The Basic Behavior of Regulatory Commissions." *A.E.R. Papers and Proc.* 62, no. 2 (May 1972): 47–54.

Levin, Richard C. "Allocation in Surface Freight Transportation: Does Rate Regulation Matter?" *Bell J. Econ.* 9 (Spring 1978): 18–45.

Locklin, D. Philip. *Economics of Transportation.* 7th ed. Homewood, Ill.: Irwin, 1972.

MacAvoy, Paul W., and Sloss, James. *Regulation of Transport Innovation: The ICC and Unit Coal Trains to the East Coast.* New York: Random House, 1967.

Meyer, J. R.; Peck, M.; Stenason, J.; and Zwick, C. *The Economics of Competition in the Transportation Industries.* Cambridge, Mass.: Harvard Univ. Press, 1959.

Rocky Mountain Motor Freight Bureau. "All-Bureau Continuing Traffic Study for 1972." Computer tape, Rocky Mountain Motor Freight Bur., Denver, 1972.

Spady, Richard H. *Econometric Estimation of Cost Functions for the Regulated Transportation Industries.* New York: Garland, 1979.

Suits, D. B.; Mason, A.; and Chan, L. "Spline Functions Fitted by Standard Regression Methods." *Rev. Econ. and Statis.* 60, no. 1 (February 1978): 132–39.

Williams, Winston. "Trucking Industry Mobilizes to Fight Rate Deregulations." *New York Times* (January 2, 1978), p. 32.

U.S. Interstate Commerce Commission. "One Percent Waybill Sample of Carload Terminations for 1972." Computer tape, ICC, Washington, 1972.

Transaction Costs, Order Placement Strategy, and Existence of the Bid-Ask Spread

Kalman J. Cohen and Steven F. Maier

Duke University

Robert A. Schwartz

New York University

David K. Whitcomb

Rutgers University

By considering investor order placement strategy, this paper demonstrates that transaction costs cause bid-ask spreads to be an equilibrium property of asset markets. With transaction costs, the probability of a limit order executing does not go to unity as the order is placed infinitesimally close to a counterpart market quote; thus, with certainty of execution at the counterpart market quote, a "gravitational pull" is generated that keeps counterpart quotes from being placed infinitesimally close to each other. An equilibrium spread is defined and its size linked to market thinness; implications are noted for the design of a trading system.

I. Introduction

This paper establishes that transaction costs in secondary asset markets cause individual investors to use order placement strategies that

We thank Amir Barnea, Avraham Beja, Richard Burton, Mark Eaker, Dan Galai, Kenneth Garbade, Thomas Ho, Wesley Magat, David Peterson, and Marshall Sarnat for their helpful comments. Earlier drafts of this paper have been presented at seminars in economics and finance at Duke University, the Hebrew University at Jerusalem, New York University, Tel-Aviv University, and the Joint National Meetings of TIMS/ORSA (New York City, May 1978).

[*Journal of Political Economy*, 1981, vol. 89, no. 2]

result in a nontrivial market bid-ask spread.[1] We define an *equilibrium* market spread and demonstrate that it will be greater for thinner securities.

The analysis fits into a growing body of literature which increasingly is being referred to as the microstructure of security markets. Stigler (1964), Demsetz (1968), West and Tinic (1971), Tinic (1972), and Tinic and West (1972, 1974), along with Farrar, Smidt, Stoll, and others involved in the institutional investor study (see U.S. Securities and Exchange Commission 1971), were among the first to focus rigorous analytical attention on the operations of security markets. More recently, microstructure theory has been explicitly considered by, among others, Garman, who coined the term (1976), Beja and Hakansson (1977), and Cohen, Maier, Schwartz, and Whitcomb (hereinafter CMSW) (1978). The analytical issues addressed involve the interplay among market participants, trading mechanisms, and the dynamic behavior of security prices. In the present paper, we study the formulation of optimal investor trading strategies and how these interact with one aspect of the dynamic behavior of security prices, the bid-ask spread. Spreads are of concern to investors because they are a variable cost of trading by market order and because they cause an inflation of transaction-to-transaction returns variance.

The pioneering analysis of bid-ask spreads was provided by Demsetz (1968). Further studies include: Tinic (1972), Tinic and West (1972, 1974), Benston and Hagerman (1974), Garman (1976), Hamilton (1976, 1978), Branch and Freed (1977), Stoll (1978*a*, 1978*b*), Ho and Stoll (1979, 1980), Newton and Quandt (1979), Schleef and Mildenstein (1979), Smidt (1979), and Amihud and Mendelson (1980). In CMSW (1979) we discuss this literature and contrast it with the formulation presented here. Except for Ho and Stoll (1980), there has been no explicit consideration of the transition from individual spreads to the market spread. It will be clear from our analysis that this transition is not a simple aggregation process and that the market spread is the product of a dynamic interaction involving many market participants. Also, previous theoretical models have generally assumed that investors can be dichotomized into two groups—immediacy demanders and immediacy suppliers. Our model of investor order placement strategy suggests that such a dichotomy will not be observed in the marketplace.

For the market, the spread is the difference between the lowest ask and the highest bid of all participants. In markets composed of many

[1] We have elsewhere (Cohen et al. 1980) considered how the impact of transaction costs on stock price movements introduces serial correlation in returns data and causes estimates of the market model beta coefficient to be biased.

traders with heterogeneous beliefs and trading propensities, one might expect to have orders at virtually every permissible price in the neighborhood of equilibrium and hence to find no significant market spread.[2] However, we show that even when expectations and trading propensities are heterogeneous, the spread is a property of asset markets that have temporarily cleared.[3] The analysis yields an existence proof of a noninfinitesimal spread with continuous pricing; a fortiori this proves for discrete prices the existence of a positive spread for nontrivial reasons.[4]

The essence of our argument is as follows. At any point in time, any investor might alternatively seek to trade via a limit order (be an immediacy supplier), trade with certainty via a market order (be an immediacy demander), or not seek to trade at all. Limit orders create the book, and market orders clear out limit orders. Because execution via market order is certain, while execution via limit order is not, it never pays for any investor to place a limit order (e.g., a bid) at a price too close to that of a counterpart limit order (e.g., an ask). Intuitively stated, as a trader contemplates placing a bid closer and closer to an ask already established on the market, he is increasingly attracted by this counterpart offer; at some point, the "gravitational pull" exerted by the established ask will dominate. The trader will "jump" his price and execute with certainty via a market order.

Section II establishes the scenario for our analysis. Section III focuses on the probability of a limit order executing and shows that, with transaction costs, this probability does not go to unity as the price at which a limit order is placed becomes infinitesimally close to a counterpart market quote. This demonstration underscores an investor's need for an order placement strategy and provides the foundation for our gravitational pull model. Section IV models the investor's order placement decision and develops conditions under which he will transmit limit or, alternatively, market orders to the market, or do nothing. The analysis is developed in a dynamic programming framework, although we are interested in the descriptive modeling of

[2] When an asset market has cleared, there is neither excess demand nor excess supply in the sense that at that moment no market participant is willing to buy the asset at a price equal to or greater than the ask, and no one is willing to sell at a price equal to or less than the bid. Hence we must have a market spread that is at least equal to the minimum allowable price change for the asset in question (on the major U.S. stock exchanges, this is 1/8 of a dollar for most common stocks). However, it is not obvious why spreads greater than minimum allowable price changes are commonly observed.

[3] While the analysis presented here is applicable to any asset market, our formal model treats the secondary market for financial securities. Security markets have two convenient properties: All units of an asset are identical, and such markets are impersonal (which means that bargaining, as distinct from trading, strategies need not be considered).

[4] See n. 2 above.

an investor's decision process rather than in actually generating a normative solution. Section V demonstrates that implementation of the strategic order placement decision (which implies the gravitational pull effect) causes a noninfinitesimal bid-ask spread to exist. This section also defines an equilibrium bid-ask spread, discusses conditions under which it will exist, and shows that it is positively related to market thinness. Section VI considers implications for the design of a security market trading system and summarizes our analysis.

II. The Scenario

Consider an investor who maximizes the expected utility of terminal wealth and, for simplicity, let him allocate funds between only two assets: a risk asset and cash (which we take to be the numeraire asset). In the absence of transaction costs, the market would be monitored continuously and appropriate transactions would be made with each change in the market price and the investor's demand propensities. Then, if price were continuous, there would be no spread and the market price would be determined by a straightforward aggregation of individual demand propensities.

However, a variety of transaction costs impact on the investor's trading decisions. The fixed (with respect to number of shares traded) costs of assessing information, monitoring the market, and conveying orders to the market imply that the investor will make trading decisions only periodically. Further, when decisions are made, he will not convey his full set of demand propensities to the market. For one thing, trades that involve sufficiently small portfolio adjustments would not justify the transaction costs incurred.[5] Also, attempts to transmit several limit orders simultaneously would be likely to overload our current system. Furthermore, a continuous auction which does not generate a Walrasian solution cannot readily handle multiple buy-sell orders that, ex ante, are intended to be alternatives.[6]

In light of transaction costs (and also taking account of the timing and magnitude of exogenous cash flows), the investor will establish a discrete set of decision points. In the analysis presented in Section IV, we take the frequency of these points as predetermined. Upon reaching a decision point, an investor can do nothing, or hit an

[5] It can be readily demonstrated that variable as well as fixed transaction costs make sufficiently small portfolio adjustments prohibitively expensive.

[6] That is, if any array of buy and sell orders from one investor is executed sequentially, the desired quantities at each price should be dependent on the exact sequence of purchases and sales followed. However, the investor does not know ex ante which specific sequence will occur. This problem could, of course, be handled (at a cost) by conditioning orders on the sequence of prior transactions (i.e., if the limit order to buy 200 at $50 executes, then sell 100 at $56).

existing limit order with his own market order, or place his own limit order at a "better" price and run the risk of its not executing. A concurrent strategy issue also exists when the investor finds it prohibitively expensive to convey his entire set of demand propensities to the market; for all intents and purposes, he must select the single best alternative to transmit.

III. The Probability of a Limit Order Executing

In this section, we establish the conditions under which the probability of execution does not approach a limit of unity as the price at which the limit order is placed is taken to be infinitesimally close to the counterpart market quote. Under these conditions we obtain a probability jump (at the counterpart market quote) that underlies the gravitational pull effect developed in Section V. By showing that the probability jump can be attributed to the existence of transaction costs, we establish the basic linkage between spreads and transaction costs.

Consider the case where an investor contemplates submitting a limit bid at the price P_t^{LB} at time t, and let price be a continuous variable. A similar argument can be constructed for the case of a limit ask. Assume that if the limit order is unfilled by the next decision point at time $t + 1$ the order will be canceled. Let L be the length of time between decision points t and $t + 1$.

Let P_t^{MA} be the market ask price at time t. Consistent with the random walk version of the efficient markets hypothesis, we make the Markov assumption that each subsequent market ask depends only on the last previous market ask. If we also assume that each change in the natural log of the market ask, Z_i, is a random variable that is independently and identically distributed over time with mean zero and variance var (Z_i), then we can model the market ask price generation process as a compound Poisson stochastic process:[7]

[7] For expositional simplicity we assume that the stochastic processes considered in this section have no drift (that is, their expected value is zero). It should be noted that even though ln P_t^{MA} is assumed to have no drift, the price series itself, P_t^{MA}, may have drift. For example, when ln (P_{t+1}^{MA}/P_t^{MA}) is normal, with mean zero and standard deviation σ, then P_{t+1}^{MA}/P_t^{MA} is log normal with mean exp $(\frac{1}{2}\,\sigma^2)$.

There will be a realization of the random variable Z_i at each point of time when any one of the following events occurs to affect the specific limit order which sets the market ask: (*a*) it is withdrawn; (*b*) it executes against a crossing buy order; or (*c*) it remains on the book but is no longer the market ask since a lower limit sell has been submitted. Note that events of types *a* and *b* necessarily result in a change in the market ask when price is continuous and utility functions are heterogeneous (which implies a zero probability of having two or more orders at a specific price). Also note that only events of type *b* are associated with transactions; hence the number of Z_i which materialize during any interval will usually exceed the number of transactions in that interval.

$$\ln P_t^{MA}(\Delta) = \ln P_t^{MA} + \sum_{i=1}^{N(\Delta)} Z_i, \tag{1}$$

where Δ is the time from the last decision point. When Δ equals L, we have $P_t^{MA}(\Delta) = P_{t+1}^{MA}$; when Δ equals $2L$, we have $P_t^{MA}(\Delta) = P_{t+2}^{MA}$, etc. The number of changes in the market ask that take place in the time interval Δ is $N(\Delta)$. We assume $N(\Delta)$ follows a Poisson process with arrival rate ν.

The next step is to determine the probability of execution (during a time interval of length L) of a limit bid which is submitted at a price P_t^{LB} greater than the current market bid (P_t^{MB}) but less than the current market ask (P_t^{MA}). Clearly, as the limit bid price approaches (from below) the market ask, the probability of execution increases. One might suppose that this probability approaches unity as the limit bid approaches the market ask; however, this need not be the case. In the Appendix we restate formally and prove:

PROPOSITION 1: If P_t^{MA} is generated by the compound Poisson process of equation (1), then no matter how close the limit bid approaches (from below) the market ask, the probability of the limit order executing is less than unity in any time interval of finite length.

Since a market order will always execute with probability one, proposition 1 gives a probability jump at the market ask.

Transaction costs are crucial to the existence of the probability jump. In the absence of transaction costs, one might expect that, following the work of Merton (1973), the logarithm of the market ask would best be described by a Wiener process with zero drift:

$$d \ln P_t^{MA}(\Delta) = \sigma dZ(\Delta), \tag{2}$$

where σ is the instantaneous variance of the process and $dZ(\Delta)$ is a standardized Wiener process. In this case, the price $P_t^{MA}(\Delta)$ would experience an infinite number of adjustments in the interval $0 \leqslant \Delta \leqslant L$. In the Appendix we restate formally and prove:

PROPOSITION 2: If $P_t^{MA}(\Delta)$ is generated by the Wiener process of equation (2), then as the limit bid approaches (from below) the market ask, the probability of the limit order executing approaches unity for all time intervals of the finite length.

Proposition 2 implies that for the Wiener process there will be no probability jump at the market ask.

We next consider whether there is a relationship between the compound Poisson process of equation (1) and the Wiener process of equation (2). In the Appendix we restate formally and prove:

PROPOSITION 3: If the random variable Z_i in the compound Poisson process of equation (1) has two equally likely possible values, $+\alpha$ and $-\alpha$, and if the arrival rate ν of this compound Poisson process increases without bound while α simultaneously decreases in such a way that $\nu\alpha^2$ is constant, then the compound Poisson process approaches the Wiener process described by equation (2) with the variance of the Wiener process, σ^2, equaling $\nu\alpha^2$.

Proposition 3 states that the compound Poisson process can be expected to approach the Wiener process under appropriate assumptions.[8] Thus the probability jump of proposition 1 would disappear as ν increased and var (Z_i) decreased. In the Appendix we restate formally and prove:

PROPOSITION 4: As the arrival rate ν in the compound Poisson process of equation (1) increases, the probability of a limit order executing increases for all $P_t^{LB} < P_t^{MA}$.

Hence the probability of a limit order executing increases at all $P_t^{LB} < P_t^{MA}$ as the activity proxy ν increases. Stated conversely, the "thinner" a security, that is, the less active are investors in submitting orders to trade in the security, the lower will be the probability of execution at each $P_t^{LB} < P_t^{MA}$, and therefore the greater will be the probability jump at P_t^{MA}.

The four propositions stated in this section (and proved in the Appendix) have important implications for the analysis of the impact of transaction costs on bid-ask spreads. Without transaction costs, the market price could be expected to behave as a Wiener process; that is, there would be an infinite number of infinitesimally small price adjustments. The investor who was considering placing a limit order could reduce the probability of his order not executing to as close to zero as desired simply by placing his order close enough to the counterpart market quotation.

On the other hand, with transaction costs and a finite number of investors, the market price would generally not behave as a Wiener process. Investors would find continuous adjustments in their portfolios too expensive, and market prices would behave as a stochastic jump process (proposition 3 shows that this process could generally be expected to approach a Wiener process as the order arrival rate is increased). One such jump process is the compound Poisson, and it also is consistent with a (martingale) efficient market.

[8] The particular distribution chosen for Z_i in proposition 3 is not critical. Other appropriately chosen discrete or continuous distributions can also be shown in the limit to go to the Wiener process.

Proposition 1 demonstrates that for such a stochastic jump process a probability jump exists. Proposition 4 has shown that the probability jump will be greater for thinner securities. In Section V we will show that the larger probability jump for thinner issues leads to equilibrium market spreads which are larger for thinner issues.

IV. A Model of Investor Order Placement Strategy

We now consider the question of when an investor will choose to trade via a market order or limit order, or not seek to trade at all. The problem is structured as follows. Because of the costs of monitoring the market, let the investor consider rebalancing his portfolio only at preselected points in time, $t = 1, 2, \ldots, T - 1$, where T is the investment horizon.[9] In order to simplify the analysis, we now consider the placement of only purchase orders for the risk asset (omitting the symmetric case of sell orders without loss of generality).

At any of the $T - 1$ decision points, the investor is faced with three possible courses of action:

a) Submit a market order to buy shares at the current market ask price of P_t^{MA}.

b) Submit a limit order to buy shares at a limit bid price of $P_t^{LB} < P_t^{MA}$.

c) Do nothing.

In modeling the investor's choice among these three alternatives, we find it convenient to make the following assumptions that simplify the analysis without materially affecting the nature of the conclusions. We assume that all orders are for a fixed number of shares ΔN and that when any market or limit order is executed, it is satisfied fully at the stated price; this avoids the tedium of writing (average) transaction price and transaction costs as functions of the number of shares exchanged and of defining probabilities of partial execution. We assume that unfilled limit orders are canceled prior to the next decision point; this avoids both the need to include additional state variables (the price and quantity of any limit order outstanding) and the need to analyze additional courses of action (submit a market order and leave the old limit order outstanding, or submit a market order and remove the limit order, etc.). Finally, we assume no lags in the transmission of information and orders; this avoids the complexity of dealing with changes in the current market quotes during the time the investor formulates and implements his decision.

[9] Note that the rebalancing points need not be the same for all investors, so that trades can occur at any time when the market is open even if specific traders are not continually in the market.

If option *a* is chosen, the market order will be executed, and the investor's holdings at time $t + 1$ become $N_{t+1} = N_t + \Delta N$, for which the investor pays a total cost of $\Delta N \cdot P_t^{MA} + C^M$ where C^M is the total cost of transmitting and executing the market order.

If option *b* is chosen, then one of two events can take place: (*b*1) The limit order is executed. The investor's share holdings then become $N_{t+1} = N_t + \Delta N$, for which the investor pays a total cost of $\Delta N \cdot P_t^{LB} + C^{L1} + C^{L2}$ where C^{L1} is the cost of transmitting a limit order to the market and C^{L2} is the cost of executing a limit order. (*b*2) The limit order does not execute and is canceled prior to the next decision point. The investor's share holdings then remain unchanged ($N_{t+1} = N_t$) and his cash is decreased by the cost of transmitting the limit order, C^{L1}. Option *b* will be chosen over *a* if the gain associated with the possibility of trading at a more favorable price outweighs the loss associated with the probability of not trading at all.

The investor must consider four subjective probability distributions in order to make an optimal decision. These are: (1) the joint probability distribution of market bid and ask prices at time $t + 1$, conditional upon the quotes at time t; (2) the probability of a limit bid order submitted at time t executing before time $t + 1$; (3) and (4) the joint probability distribution of market bid and ask prices at time $t + 1$, conditional upon the quotes at time t and further conditional on whether a limit bid submitted at time t either did or else did not execute prior to time $t + 1$. In the Appendix we discuss in more detail these four subjective probability distributions (only three of which are independent).

A dynamic programming model of the investor's choice among options *a*, *b*, and *c* is formulated in the Appendix. The investor is assumed to maximize his expected utility of terminal wealth, max (U_1, U_2, U_3), where U_1 is the expected terminal utility of choosing option *a*, trading via a market order; U_2 is the expected terminal utility of choosing option *b*, seeking to trade via a limit order; and U_3 is the expected terminal utility of choosing option *c*, doing nothing. It is convenient to focus on the utility gain, ΔU_1 or ΔU_2, which results from choosing option *a* or *b* rather than option *c*: $\Delta U_1 = U_1 - U_3$; $\Delta U_2(P_t^{LB}) = U_2(P_t^{LB}) - U_3$ (note that U_2 and ΔU_2 are functions of the limit bid price). Clearly $\Delta U_2(P_t^{LB} = P_t^{MA}) = \Delta U_1$, since, at this price, the market order and limit order strategies are effectively the same (the probability is unity of the limit order executing at a price of P_t^{MA}).

Let us now consider the conditions under which $U_1 > \max(U_2, U_3)$, in which case the investor will submit a market order, or $U_2 > \max(U_1, U_3)$, in which case the investor will submit a limit order. Suppose that at the current market quotes the do-nothing strategy is dominated (i.e., $\max[U_1, U_2] > U_3$). Given our utility gains ΔU_1 and

$\Delta U_2(P_t^{LB})$ and the probability function for limit order execution developed in Section III above, we know the utility gain from placing a market order and can readily obtain an expected utility gain function for the limit order strategy. These are depicted in figure 1; the shape of the function is explained as follows:

1. While the utility of a consummated trade decreases monotonically with P_t^{LB}, the probability of execution increases with P_t^{LB}, with two probability jumps (at the market bid and the market ask). The jump at P_t^{MB} simply reflects the institutional reality that orders placed at prices less than or equal to P_t^{MB} would not have priority over the current market bid, whereas an order placed at a price above P_t^{MB} would be at the top of the limit order book. The jump at P_t^{MA} follows from proposition 1.

2. Since the probability of execution is constant to the right of P_t^{MA}, the expected utility gain has a peak at P_t^{MA}, corresponding to the utility gain of transacting by a market order.

3. The probability of execution increases between P_t^{MB} and P_t^{MA}, with the greatest relative increase just to the right of P_t^{MB}. We expect this large probability increase in this neighborhood because the strategy considerations of other investors might lead them to place limit orders just to the right of P_t^{MB} in order to capture the largest price advantage. Hence the second peak in the expected utility gain function will occur at some point P'', between P_t^{MB} and P_t^{MA}.

4. The probability of execution decreases rapidly in a neighborhood just to the left of P_t^{MB} because of existing limit orders on the book. However, since there would be a clustering of limit orders near

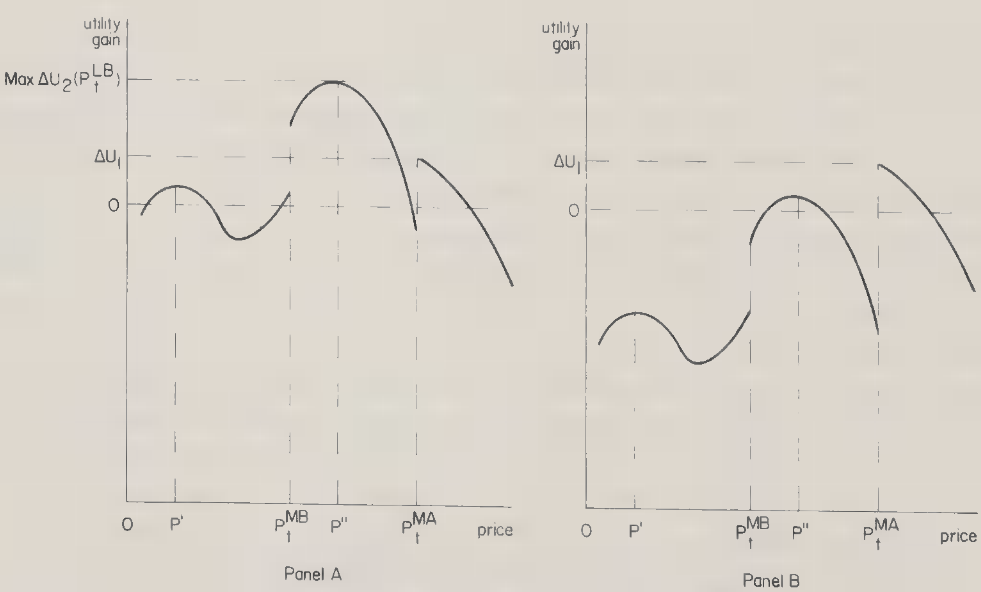


FIG 1.—Illustrative utility gain functions

P_t^{MB}, we would expect the probability of execution to decrease more slowly as the limit price, P_t^{LB}, moves further from P_t^{MB}. Hence, we might find a third peak in the expected utility gain function to occur at some point P' to the left of P_t^{MB}.

Which of the three peaks in $\Delta U_2(P_t^{LB})$ will dominate depends on the particular form of the probability function for limit order execution as well as the expected utility function. In the particular diagram presented in figure 1*A*, $\Delta U_2(P_t^{LB})$ has positive values (hence a do-nothing strategy is dominated), has values greater than ΔU_1 (hence the market order strategy is dominated), and gives a global maximum at P''. Assuming only one limit order per investor, the optimal strategy is to place a limit bid at P'', and hence P'' will become the new market bid. The setting of such a bid quotation can be considered a Stigler (1961)–type search activity.[10]

Clearly the limit order strategy need not always be superior; in selecting this strategy, the investor also accepts the chance that the limit order may fail to execute, in which case the investor would be worse off than had he done nothing since he would have also lost the cost of placing the limit order, C^{L1}. If the probability of failing to execute is high enough and $\Delta U_1 > 0$, the investor will prefer the market order strategy as illustrated in figure 1*B*.[11]

V. The Market Spread

It is clear from the preceding analysis (Sec. IV) that, with a continuous auction market, each investor's order placement decision is made with reference to prices already established on the market—the bid and ask quotations which define the market spread. In turn, each investor's order may affect the market spread to which subsequent traders react. Hence, the market spread is the product of a dynamic interactive process. In this section we show that a nonzero market spread must exist, define the equilibrium market spread, and relate a security's equilibrium spread to its thinness.

The limit order book comprises the limit orders transmitted to the market by a subset of the many traders in a security, and the spread is essentially a gap in the limit order book. Therefore, having estab-

[10] By the placement of such a limit order, a seller or buyer announces his propensity to trade with the hope of getting the attention of a counterpart market participant who would also be willing to trade.

[11] There are two other versions of fig. 1. One would be analogous to panel *A*, but with the global maximum at P' rather than at P''; in this case, the investor would submit a limit bid at P' (below the market bid). The second version would be where the zero point on the utility gain axis is higher than both max $\Delta U_2(P_t^{LB})$ and ΔU_1; in this case, the investor would do nothing.

lished that individual investors will sometimes seek to trade via limit orders, we must show why a noninfinitesimal gap in the array of such orders exists between the market quotes. We do so by considering the impact that the jump in the probability of a limit order executing (Sec. III) has on the investor's optimal order placement strategy (Sec. IV).

We have established the conditions under which a limit order will be placed; we can now consider whether or not limit orders will necessarily be placed so as always to preserve a nonzero bid-ask spread. By continuity of the utility function, and because of the discontinuous drop in the probability of execution that must occur at the market ask as we move to lower prices, we must have a discontinuous drop in $\Delta U_2(P_t^{LB})$ as we move past the market ask to lower prices. Then, because of continuity of the probability function until we get to the market bid (at which point another discontinuous drop occurs), it necessarily follows that $\Delta U_2(P_t^{LB})$ cannot exceed ΔU_1 within a nonzero neighborhood to the left of P_t^{MA}. Hence, no limit bid will be placed infinitesimally close to the market ask, and therefore we must have a nonzero bid-ask spread.

Intuitively viewed, consider an investor who is thinking about whether or not to place a limit bid at some price below the market ask. As the potential limit price rises toward the market ask, it becomes relatively more attractive to transact at the market ask with certainty. At some point, the investor will prefer buying at a discretely higher, unitary probability of execution. In other words, certainty of execution at the market ask creates a kind of gravitational pull which causes investors to jump their price once their potential limit bid gets close enough to the market ask. This clears out limit bids when they get too close to the market ask, thereby leaving a nonzero spread between the market bid and ask.

We next consider the issue of an equilibrium market spread.

Definition: In a dynamic trading process we define the equilibrium market spread as the bid-ask spread at which, for the next instant of time, the probability of the spread increasing is equal to the probability of the spread decreasing.[12]

This definition does not imply that if away from equilibrium the

[12] Note that the probability is not conditioned on the current status of individual investors, recent prices, or other market conditions.

Note that we use the term "equilibrium market spread" to refer to one particular point in the probability distribution of the market spread, rather than to refer to a statistic for the central tendency of the entire distribution. We do not adopt a central tendency definition (such as expected value, mode, or median) of the equilibrium spread because it does not provide a ready link to the market forces that generate the spread and to thinness. For empirical research, the average market spread is likely to be a reasonable proxy for our definition of the equilibrium market spread.

spread will necessarily move toward it on the next order, but only that it is more likely to do so than not.

We assume that a unique equilibrium exists. To examine the reasonableness of this assumption, first consider an arbitrarily small spread. Because the probability of a limit order being placed has a discrete jump at P_t^{MA}, we have already argued that there will be no limit orders placed in some nonzero neighborhood below P_t^{MA} by *all* investors seeking to buy the security. (The generalization from a single investor to all investors takes place simply by choosing the smallest neighborhood found among all investors.) Therefore, for a sufficiently small bid-ask spread, the unconditional probability of the spread increasing must exceed the unconditional probability of the spread decreasing as analyzed from the standpoint of the buyer. A similar argument holds for the seller.

Now consider what occurs when the spread widens. The larger the spread is, the greater the potential utility gain from an optimally priced limit order, while the utility gain from a market order will remain unchanged or fall. The gain in utility occurs because the shares can be bought (sold) at a lower (higher) price via limit order than by market order.[13] With the rise in utility $U_2(P_t^{LB})$, investors will begin to shift their preferences from market orders to limit orders. As more and more investors shift to limit orders, the probability of the spread increasing falls, while the probability of it decreasing rises.

One of two situations must occur. Either the spread would tend to reach a point at which the two probabilities would be equal, in which case it would be the equilibrium market spread, or the limit order book would be empty on one or both sides.[14] Since when the order book is empty market orders cannot transact, we would be at a point where the probability of a limit order is greater than that of a market order. Although such a point might be achieved without the equilibrium market spread ever being attained, we believe it would be atypical.[15]

[13] We assume that the spread does not impart any information to the investor on the direction of future price movements.

[14] No other possibility exists. So long as there is a greater probability that the spread will increase than that it will decrease, the spread will on expectation grow—limited only by the collapse of the limit order book on one side or the other.

[15] With regard to the assumption of uniqueness, investors might have utility functions which could lead to more than one equilibrium market spread. Consider the case where an investor views a very wide spread as signaling the advent of new information which may cause a price adjustment of unknown direction. The investor sensing greater risk might now choose a do-nothing strategy over placing either a limit or market order. This could support a very wide spread as a new market equilibrium spread. Of course, investors would eventually conclude that the new information was not forthcoming, and trading would resume at its former frequency with the old equilibrium spread reestablished.

Our definition of the equilibrium market spread provides a direct link between the market spread and thinness. Recall that proposition 4 showed that the probability of a limit order executing decreases, for all values of P_t^{LB}, as a security's order arrival rate decreases. Thus, ceteris paribus, thin issues have a lower probability of limit orders executing than do thick issues. This lower probability in turn decreases the proportion of investors choosing limit orders over market orders at any given size of the spread; this implies that, for a thinner issue, a wider spread would be required for the forces that increase and decrease the spread to be in equilibrium. Hence we have:

PROPOSITION 5: Thinner securities will, ceteris paribus, have larger equilibrium market spreads.

VI. Conclusion

We have presented a proof of the existence of market spreads in markets with many traders. The formulation treats continuous prices and allows for heterogeneous trading propensities. The proof is not dependent on the demand of traders for immediacy or on the cost to market makers of providing immediacy.

The literature on bid-ask spreads does not appear to have recognized that aggregation from individual to market spreads is a considerably more complex process than the standard aggregation from individual to market demand and supply functions. Neither has it been established that ordinary investors may sometimes seek to trade via limit orders (and at other times via market orders), hence that these investors will sometimes supply (and at other times demand) immediacy, and that in choosing between market and limit orders investors implement an order placement strategy. In addressing these issues, we have sought to establish the links between transaction costs, individual investor order placement strategy, market thinness, and market spreads.

We have first established that, with transaction costs, the probability of a limit order executing does not rise to unity as the price at which the order is placed gets infinitesimally close to a counterpart market quote. We have next shown that the resulting investor trading strategies generate what we have referred to as a gravitational pull effect. Essentially, in the neighborhood of the current market bid and ask quotations, what might otherwise have been limit orders are instead submitted as market orders (at slightly less desirable prices) so as to achieve certainty of execution. These market orders trigger trades which clear limit orders off the book, widening the market spread. The gravitational pull effect explains why market spreads may be

substantial even in markets composed of many traders. Finally, we have defined an equilibrium market spread (where the forces that tend to widen and to narrow the spread are in balance) and have shown it to be positively related to a security's thinness (measured inversely by the order arrival rate).

Our formulation has several implications for the design of a market system. A primary objective of system design should be to expand the extent and frequency with which investors interact with the market by minimizing various transaction costs. Decreasing variable transaction costs will decrease individual spreads and generate a greater order flow; decreasing the costs of monitoring and communicating with the market will also increase the frequency with which investors rebalance their portfolios; and consolidating the currently fragmented system (by, e.g., instituting a consolidated limit order book) will reduce search costs and further shrink spreads by increasing the effective thickness of the market. These costs are all a function of market structure and hence should be amenable to reduction by appropriate system design. However, a major cost of interacting with the market is the cost of decision making, and this might not be subject to significant reduction by exchange organization. For this reason, it is possible that, especially for thinner issues, spreads will remain sizable in a restructured national market system.

Appendix

In this Appendix, we restate formally and prove propositions 1–4 of Section III above and develop the more technical aspects of the arguments presented in Section IV.

Section III

Equation (1) of Section III presents a compound Poisson process as the stochastic process which generates a sequence of market ask prices over time. This can be used to determine the probability that a limit bid order will execute during a particular time period.

Suppose that the price of the limit order to be submitted is greater than the current market bid but less than the current market ask, that is, that $P_t^{MB} < P_t^{LB} < P_t^{MA}$. To determine the probability that the potential limit bid will execute in a time interval of length L, we must find the probability that $P_t^{MA}(\Delta)$ will decrease to a price equal to or less than P_t^{LB}.[16] Therefore, let $Y(L, P_t^{MA}, P_t^{LB})$ be the probability that the minimum value that $P_t^{MA}(\Delta)$ achieves in the interval $0 \leq \Delta \leq L$ is equal to or less than P_t^{LB}. As P_t^{LB} approaches P_t^{MA}, we would expect the probability of the potential limit bid executing to increase, since the amount that $P_t^{MA}(\Delta)$ would have to decline would be reduced. However, in the

[16] For P_t^{MA} discretely greater than P_t^{LB}, the possibility of another investor submitting a limit order with a price greater than P_t^{LB} will only tend to decrease the probability of execution of the original limit order. Thus our proof of existence of the probability jump at P_t^{MA} is conservative.

limit as P_t^{LB} approaches P_t^{MA} from below, without further analysis it is unclear how far the probability will rise. Therefore, let

$$\phi(L, P_t^{MA}) = 1 - \lim_{P_t^{LB} \to P_t^{MA}} Y(L, P_t^{MA}, P_t^{LB}),$$

where the limit is understood to be from below. We now prove:

PROPOSITION 1: If $P_t^{MA}(\Delta)$ is generated by the compound Poisson process of equation (1), then $\phi(L, P_t^{MA}) > 0$ for all intervals of length $L < \infty$.[17]

PROOF: Since Z_i is stochastic with mean zero, $P(Z_i \geq 0) > 0$. Furthermore, the value of the Poisson random variable $N(L)$ will be finite in any interval of length L less than infinity. Therefore, the probability of $N(L)$ consecutive Z_i observations that are greater than or equal to zero is given by $[P(Z_i \geq 0)]^{N(L)}$, which, since $N(L)$ is finite, must be strictly greater than zero. Notice that if all the Z_i observations are greater than or equal to zero, the value of $P_t^{MA}(\Delta)$ must be greater than or equal to P_t^{MA} throughout the interval $0 \leq \Delta \leq L$; thus the limit order P_t^{LB} would not have executed. This is sufficient to demonstrate the probability jump. Clearly, there are other sample paths that $P_t^{MA}(\Delta)$ could have followed which also would have failed to execute the limit order.

PROPOSITION 2: If $P_t^{MA}(\Delta)$ is generated by the Wiener process of equation (2), then $\phi(L, P_t^{MA}) = 0$ for all intervals of length $L < \infty$.

PROOF: By the reflection principle for a continuous Wiener process (see Karlin 1968, pp. 276–77) and since $\ln P_t^{MA}(\Delta)$ is driftless,

$$\begin{aligned} Y(L, P_t^{MA}, P_t^{LB}) &= \Pr\{\min_{0\leq\Delta\leq L} P_t^{MA}(\Delta) \leq P_t^{LB}\} = \Pr\{\min_{0\leq\Delta\leq L} \ln P_t^{MA}(\Delta) \leq \ln P_t^{LB}\} \\ &= 2\Pr\{\ln P_{t+1}^{MA} < \ln P_t^{LB}\} = 2\Pr\{P_{t+1}^{MA} < P_t^{LB}\} \\ &= \frac{2}{\sigma\sqrt{2\pi L}} \int_{-\infty}^{P_t^{LB}} \exp\left\{-\frac{1}{2L}\left(\frac{x - P_t^{MA}}{\sigma}\right)^2\right\} dx, \end{aligned}$$

where the latter probability distribution follows from the definition of the Wiener process. Now

$$\lim_{P_t^{LB} \to P_t^{MA}} Y(L, P_t^{MA}, P_t^{LB}) = \frac{2}{\sigma\sqrt{2\pi L}} \int_{-\infty}^{P_t^{MA}} \exp\left\{-\frac{1}{2L}\left(\frac{x - P_t^{MA}}{\sigma}\right)^2\right\} dx.$$

Substituting the variable $y = (x - P_t^{MA})/(\sigma\sqrt{L})$, the preceding limit equals $(2/\sqrt{2\pi}) \int_{-\infty}^{0} \exp\{-\frac{1}{2} y^2\} dy = 2(\frac{1}{2}) = 1$.

PROPOSITION 3: If the random variable Z_i is expressed as a Bernoulli random variable, with $Pr(Z_i = \alpha) = Pr(Z_i = -\alpha) = 1/2$, and if the arrival rate ν of the Poisson process $N(\Delta)$ goes to infinity, while simultaneously reducing the size of α in such a way that $\nu\alpha^2$ remains constant, the compound Poisson process approaches the Wiener process described by equation (2) with $\sigma^2 = \nu\alpha^2$.

PROOF: This follows from the theorem that the characteristic function for the compound Poisson approaches that of the Wiener process (see Parzen 1962, p. 99).

PROPOSITION 4: $[\partial Y(L, P_t^{MA}, P_t^{LB})]/\partial\nu > 0$ for all $P_t^{LB} < P_t^{MA}$ and $L < \infty$.

PROOF: By the Markov property of the compound Poisson process, increasing (decreasing) the order arrival rate by some factor λ is identical to

[17] This proposition would also hold for other stochastic jump processes where the number of price changes is finite in any finite interval.

increasing (decreasing) the length of time L between decision points by λ. When λ is increased, those sample paths of the process which initially satisfied

$$\min_{0 \leq \Delta \leq L} P_t^{MA}(\Delta) \leq P_t^{LB}$$

will still continue to do so. On the other hand, some sample paths where

$$\min_{0 \leq \Delta \leq L} P_t^{MA}(\Delta) > P_t^{LB}$$

will now satisfy the inequality

$$\min_{0 \leq \Delta \leq \lambda L} P_t^{MA}(\Delta) \leq P_t^{LB}.$$

Therefore, $Y(L, P_t^{MA}, P_t^{LB})$ would increase for $\lambda > 1$ for all values of $P_t^{LB} < P_t^{MA}$. Similarly, $Y(L, P_t^{MA}, P_t^{LB})$ would decrease if the order arrival rate λ is decreased.

Section IV

We now present the more technical aspects of the model of an investor's order placement strategy discussed in Section IV. The subjectively determined probability distributions are specified as follows. First, the probability distribution of future market bid and ask prices is given by $h(P_{t+1}^{MA}, P_{t+1}^{MB} \mid P_t^{MA}, P_t^{MB})$, where h is a joint density function for market asks and bids $(P_{t+1}^{MA}, P_{t+1}^{MB})$ at time $t + 1$, given the prices at t. Consistent with random walk models of security price behavior, we condition future prices only on current prices. Consistent with Section III, we let the investor's subjective probability of a limit order executing before $t + 1$ be given by $p(P_t^{LB}, P_t^{MA}, P_t^{MB})$, where P_t^{LB} is the limit bid price. (Again, this is consistent with random walk theory, since we need know only current market quotes to predict whether or not the limit order will be executed.) Last, we assume the investor determines conditional probability distributions of future prices in the event of either a successful or unsuccessful limit order. That is, for a limit bid at price P_t^{LB} submitted at time t, let $k(P_{t+1}^{MA}, P_{t+1}^{MB} \mid P_t^{MA}, P_t^{MB}, P_t^{LB})$ be the joint density function of market bid and ask prices at time $t + 1$ if the limit bid executes prior to $t + 1$, and let $l(P_{t+1}^{MA}, P_{t+1}^{MB} \mid P_t^{MA}, P_t^{MB}, P_t^{LB})$ be the joint density function at time $t + 1$ if the limit bid fails to execute prior to $t + 1$. Note that only three of the four subjective probability distributions h, p, k, and l can be independently determined, since by Bayes's theorem we must have

$$\begin{aligned} h(x, y \mid P_t^{MA}, P_t^{MB}) &= p(P_t^{LB}, P_t^{MA}, P_t^{MB}) k(x, y \mid P_t^{MA}, P_t^{MB}, P_t^{LB}) \\ &\quad + [1 - p(P_t^{LB}, P_t^{MA}, P_t^{MB})] l(x, y \mid P_t^{MA}, P_t^{MB}, P_t^{LB}) \end{aligned}$$

for all possible values of P_t^{MA}, P_t^{MB}, and P_t^{LB}.

We can now give a formal statement of the model. Given observed market quotes of P_t^{MA} and P_t^{MB} and holdings by the investor of N_t shares of the security and S_t dollars in cash, at time t the investor's expected utility of terminal wealth can be written as $\psi_t = f_t(P_t^{MA}, P_t^{MB}, N_t, S_t)$.[18]

[18] Note that the investor's underlying utility function measures the utility of the wealth he will possess at some horizon T, where $T > t$. The investor will choose those decisions at times $t, t + 1, \ldots, T - 1$ which maximize the expected utility of his terminal wealth. As of time t (the investor's current decision point), viewing the expectations operator as ranging over relevant random variables pertaining to times between t and T, we define ψ_t to be the expected utility of the investor's terminal wealth. Clearly this depends upon the assets the investor has at time t (N_t and S_t) and the current market prices at which the investor can buy or sell shares (P_t^{MA} and P_t^{MB}).

We now derive an expression for ψ_t in terms of the various probability assessments and costs. Write $\psi_t = \max(U_1, U_2, U_3)$, where U_1 = expected terminal utility if a market order is placed at time t, U_2 = expected terminal utility if a limit order is placed at time t, and U_3 = expected terminal utility of doing nothing at time t. More precisely, letting x and y be values for the ask and bid prices at time $t + 1$, respectively, we have

$$U_1 = \int_0^\infty \int_0^\infty f_{t+1}(x, y, N_t + \Delta N, S_t - C^M - N \cdot P_t^{MA}) \cdot h(x, y \mid P_t^{MA}, P_t^{MB}) dxdy.$$

As stated in Section IV, we assume that the number of shares to be purchased, ΔN, is fixed. Define the function

$$\begin{aligned} U_2(P_t^{LB}) = p(P_t^{LB}, P_t^{MA}, P_t^{MB}) \int_0^\infty \int_0^\infty & f_{t+1}(x, y, N_t + \Delta N, \\ & S_t - C^{L1} - C^{L2} - \Delta N \cdot P_t^{LB}) \cdot k(x, y \mid P_t^{MA}, P_t^{MB}, P_t^{LB}) dxdy \\ & + [1 - p(P_t^{LB}, P_t^{MA}, P_t^{MB})] \int_0^\infty \int_0^\infty f_{t+1}(x, y, N_t, S_t - C^{L1}) \\ & \cdot l(x, y \mid P_t^{MA}, P_t^{MB}, P_t^{LB}) dxdy \end{aligned}$$

for all possible prices of the limit order. Then we have

$$U_2 = \max_{P_t^{LB}} U_2(P_t^{LB}).$$

Finally, for the do-nothing option we have

$$U_3 = \int_0^\infty \int_0^\infty f_{t+1}(x, y, N_t, S_t) \cdot h(x, y \mid P_t^{MA}, P_t^{MB}) dxdy.$$

The dynamic programming recursion permits us to obtain a solution for f_t in the order $T - 1, T - 2, \ldots, 2, 1$. This is possible since we assume the utility value for f_T is known for all values of the parameter P_T^{MA}, P_T^{MB}, N_T, and S_T. We have defined this recursion in terms of the four state variables P_t^{MA}, P_t^{MB}, N_t, and S_t. We also have two decision variables, P_t^{LB} and the decision as to which of the three courses of action to take.

References

Amihud, Yakov, and Mendelson, Haim. "Dealership Market: Market-making with Inventory." *J. Financial Econ.* 8 (March 1980): 31–53.

Beja, Avraham, and Hakansson, Nils H. "Dynamic Market Processes and the Rewards to Up-to-Date Information." *J. Finance* 32 (May 1977): 291–304.

Benston, George J., and Hagerman, Robert L. "Determinants of Bid-Asked Spreads in the Over-the-Counter Market." *J. Financial Econ.* 1 (December 1974): 353–64.

Branch, Ben, and Freed, Walter. "Bid-Asked Spreads on the Amex and the Big Board." *J. Finance* 32 (March 1977): 159–63.

Cohen, Kalman J.; Hawawini, Gabriel A.; Maier, Steven F.; Schwartz, Robert A.; and Whitcomb, David K. "Implications of Microstructure Theory for Empirical Research on Stock Price Behavior." *J. Finance* 35 (May 1980): 249–57.

Cohen, Kalman J.; Maier, Steven F.; Schwartz, Robert A.; and Whitcomb,

David K. "The Returns Generation Process, Returns Variance, and the Effect of Thinness in Securities Markets." *J. Finance* 33 (March 1978): 149–67.

———. "Market Makers and the Market Spread: A Review of Recent Literature." *J. Financial and Quantitative Analysis* 14 (November 1979): 813–35.

Demsetz, Harold. "The Cost of Transacting." *Q.J.E.* 82 (February 1968): 33–53.

Garman, Mark B. "Market Microstructure." *J. Financial Econ.* 3 (June 1976): 257–75.

Hamilton, James L. "Competition, Scale Economies, and Transaction Cost in the Stock Market." *J. Financial and Quantitative Analysis* 11 (December 1976): 779–802.

———. "Marketplace Organization and Marketability: NASDAQ, the Stock Exchange, and the National Market System." *J. Finance* 33 (May 1978): 487–503.

Ho, Thomas, and Stoll, Hans R. "Optimal Dealer Pricing under Transactions and Return Uncertainty." Working Paper, New York Univ., Graduate School Bus. Admin., 1979.

———. "On Dealer Markets under Competition." *J. Finance* 35 (May 1980): 259–67.

Karlin, Samuel. *A First Course in Stochastic Processes.* 2d ed., enl. New York: Academic Press, 1968.

Merton, Robert C. "An Intertemporal Capital Asset Pricing Model." *Econometrica* 41 (September 1973): 867–87.

Newton, William, and Quandt, Richard E. "An Empirical Study of Spreads." Working Paper, Princeton Univ., Dept. Econ., 1979.

Parzen, Emanuel. *Stochastic Processes.* San Francisco: Holden-Day, 1962.

Schleef, Harald J., and Mildenstein, Eckhard. "A Dynamic Model of the Security Dealer's Bid and Ask Prices." Paper presented at meetings of the Western Economic Association, Las Vegas, 1979.

Smidt, Seymour. "Continuous vs. Intermittent Trading on Auction Markets." Working Paper, Cornell Univ., Graduate School Bus. and Public Admin., 1979.

Stigler, George J. "The Economics of Information." *J.P.E.* 69, no. 3 (June 1961): 213–25.

———. "Public Regulation of the Securities Markets." *J. Bus.* 37 (April 1964): 117–42.

Stoll, Hans R. "The Supply of Dealer Services in Securities Markets." *J. Finance* 33 (September 1978): 1133–51. (*a*)

———. "The Pricing of Security Dealer Services: An Empirical Study of NASDAQ Stocks." *J. Finance* 33 (September 1978): 1153–72. (*b*)

Tinic, Seha M. "The Economics of Liquidity Services." *Q.J.E.* 86 (February 1972): 79–93.

Tinic, Seha M., and West, Richard R. "Competition and the Pricing of Dealer Service in the Over-the-Counter Stock Market." *J. Financial and Quantitative Analysis* 7 (June 1972): 1707–27.

———. "Marketability of Common Stocks in Canada and the U.S.A.: A Comparison of Agent versus Dealer Dominated Markets." *J. Finance* 29 (June 1974): 729–46.

U.S. Securities and Exchange Commission. *Institutional Investor Study Report.* 92d Cong., 1st sess. House Document no. 92–64, 1971.

West, Richard R., and Tinic, Seha M. *The Economics of the Stock Market.* New York: Praeger, 1971.

Land Value Capitalization in Local Public Finance

David A. Starrett

Stanford University

We explore the conditions under which the welfare benefits of local public goods projects will be capitalized into land values. We find two types of sufficient conditions, one involving similar communities and the other involving differentiated communities. The form of capitalization differs between these cases, and we explore the nature of these differences. We also examine intermediate cases and identify models in which there will be no capitalization.

I. Introduction

The idea that the value of local public goods projects will be capitalized into land rents has a long tradition in the public finance literature, dating back at least as far as Tiebout (1956). Empirical evidence has been mixed, but some positive results have been reported by Oates (1969), among others. This simple idea is potentially very powerful. It provides a simple measure of project benefits which reveals (if perhaps after the fact) individual preferences for public goods. And if rent changes can be predicted in advance, it may provide an objective function for community decision making.

Furthermore, the presence of capitalization forces has important equity implications. To the extent that capitalization occurs, benefits tend to accrue to landlords at the expense of renters. We will show that capitalization reflects a real welfare increase only if these two groups are considered equally deserving. If, on the other hand, renters are considered more deserving, then the presence of capitalization clearly will worsen the distribution of benefits within society.

[*Journal of Political Economy*, 1981, vol. 89, no. 2]

The "capitalization hypothesis" is interesting from a purely theoretical point of view as well. It is clear that increases in land rent values are not themselves a net social benefit; owners of the land will benefit, but renters will suffer. Indeed, in the standard surplus theory of welfare measurement, changes in land prices (or prices in any other competitive market) never appear as net benefits (in aggregate, the gains to winners just offset losses to losers). Thus, if capitalization occurs, it must happen because the true benefits induce a corresponding change in rents.

We will argue that there are two separate and independent forces which can induce capitalization, one involving competitive forces between communities and one involving such forces within a particular community. Furthermore, these two forces have quite different implications for (1) what benefits are capitalized and (2) where those benefits are capitalized. Thus, we will derive several different capitalization theorems and explore the conditions under which each holds. In the process, we will find some conditions under which capitalization does not occur at all or occurs only partially.

Let us refer to capitalization which derives from forces between communities as "external capitalization" and capitalization which derives from forces within communities as "internal capitalization." The intuitive argument for external capitalization is quite simple. Suppose that a project is built in one community designed to make people better off there. Now, if people in all communities have similar tastes and are free to move among the communities, outsiders will necessarily be attracted to the project-building community. And they will continue to move until the welfare incentive disappears. The only factor which can stop this movement is a differential location cost, that is, an increase in land rents in the project-building community. This type of argument is the one used by Polinsky and Shavell (1976) to justify a form of capitalization. However, the Polinsky/Shavell model is not very general, since the benefits accrue to absentee landlords. We will show below that increases in land values generally will overstate true net social benefit, although they may be a good approximation to community net benefit.

Naturally, when we drop the assumption of free mobility between communities or the assumption of similar tastes between communities, the intuition for external capitalization is weaker. Indeed, several authors have argued that we should not expect land value capitalization in specialized (or Tiebout) communities which are immune to migration forces due to differences in tastes (see Hamilton 1975 and Polinsky and Rubinfeld 1978).

However, there is a quite different argument for internal capitalization which is based on specific properties of local public goods.

Presumably public goods are local in character when proximity matters. For example, one has to make trips in order to appreciate such public goods as parks, museums, civic centers, sports complexes, and even roads and highways. Now, if there is a positive correlation between the amounts of public goods and the differential desirability of various locations, some degree of capitalization will occur as rents adjust. We will show later exactly what conditions are necessary for full capitalization. This type of argument has been used in the literature to justify the use of rent gradients for measuring the cost of pollution near a source (such as an airport or factory).[1] It is also in the spirit of arguments for capitalization given by a number of authors, among them Strotz (1968), Lind (1973), and Pines and Weiss (1976). These authors essentially start with the assumption that a project will differentially improve land quality and explore whether or not those improvements will translate into rent increases. The results are instructive, but the underlying approach seems to beg the question somewhat. The primary question is, Will a local public project change the differential quality of land? For the case of irrigation projects discussed by Lind, the answer is clearly yes, but in other cases it is less obvious. Indeed, we will show that, for some types of public goods, no capitalization occurs at all.

We will want to pay attention to the system of taxation in discussing both types of capitalization. One might guess that the form and nature of capitalization will depend on whether or not the public goods are financed out of a property tax. We will show that this is sometimes, but not always, the case.

We adopt a somewhat generalized version of the Polinsky/Shavell community model in the sequel. This model has a number of special features but is chosen because it facilitates the exposition and makes for easy comparisons with the preceding literature. However, many of the results reported here remain valid in a much more general economic model. The interested reader is referred to Starrett (1978) for some general results on capitalization.

II. A Prototype Model

Throughout this section we deal with a model in which there is a single local public good (q) being produced independently by a number of communities. Further, we will suppose that projects initiated in a community will not affect private goods prices other than land rents.

[1] For an example of this argument in the literature, see Freeman 1971.

This may seem a serious restriction, but it is not, really, since changes in other prices will have the usual canceling effects on the two sides of the market. With nonland private goods prices fixed, we can aggregate all these goods into a single commodity and let this commodity be the numeraire in what follows.

Agents must engage in effort (take trips) to enjoy the public good. Naturally, such trips are costly; here we will assume that the expense can be represented by a simple cost function in terms of the numeraire: $f(g,s)$, where g is the number of trips taken and s stands for location within the community. Location will be specified by dividing the community's land into discrete zones and assuming that location can adequately be specified by designating the appropriate zone.

Each consumer gets to choose a zone and an amount of land within the zone (ℓ). The market for land will be assumed competitive, and r_s will stand for the rental rate in zone s.

Consumers earn income in terms of the numeraire. We will want to separate this income into three terms: profit shares (Π), rental income on land (R), and other income (mainly labor) (Y). Thus, an individual in income class i has income $I^i = \Pi^i + Y^i + R^i$. Such an individual will be required to pay taxes T^i. The T^i may be a function of other parameters, depending on the nature of taxation.

A consumer of preference type a has a preference function of the form $U^a[q,g,\ell,I^i - T^i - f(g,s) - r_s\ell]$. The arguments of this function are the level of public goods provision, the number of trips taken, the amount of land consumed, and the amount of other private goods consumed. It seems appropriate that both g and q should affect utility. For example, if the public good is a museum, agents will care about the size of the museum as well as the number of visits made.

Some of the results reported below will depend on the properties of U and f. Specific assumptions will be introduced as we go along, but the following general discussion may be helpful here. It seems reasonable that the marginal cost of trips ($\partial f/\partial g$) will increase as the distance from the public good increases. We expect to find that use decreases with distance if and only if this assumption holds. Also, it is reasonable to suppose that the marginal benefit from public goods ($\partial U/\partial q$) increases with use (g). This marginal benefit will decrease with distance whenever both assumptions hold.

Each consumer in a particular community gets to choose g, ℓ, and s. (At a later stage such a consumer may be able to choose a community of residence as well.) We will assume throughout that there is free mobility within the community, so that s is a free choice. It is analytically convenient to think of each resident as making a conditional

choice of g and ℓ for each s. This process generates a first-stage optimization problem:

$$\max_{g\ell} U^a[q,g,\ell,I^i - T^i - f(g,s) - r_s\ell]. \tag{1}$$

A solution to problem (1) defines demand functions $g_s^{ai} \equiv g^a(q,I^i - T^i,r_s,s)$ and $\ell_s^{ai} \equiv \ell^a(q,I^i - T^i,r_s,s)$ and an indirect utility function $V_s^{ai} \equiv V^a(q,r_s,I^i - T^i,s)$. At a later "stage," the consumer makes a discrete location choice, seeking to $\max_s V_s^{ai}$. We will use the notation V^{ai} to denote the final indirect utility function.

The remainder of the community economy consists of firms, about which we need to say little except that they behave competitively and are not affected by the public goods, and the government, which produces the public good from private goods, paying for those goods with revenues from the taxes levied on households. Letting $\Gamma(q)$ stand for the public goods cost function, the government must satisfy a balanced budget condition $T = \Gamma(q)$, where T stands for total tax revenue.

We assume that the private sector is always in market equilibrium. For nonland commodities, this simply means that markets clear at fixed (unchanging) prices. For land, it means that rental rates must adjust until the market for land clears in each zone.

We will formulate questions about capitalization as follows. We measure welfare using a standard Bergson-Samuelson formulation. We ignore distributional considerations by assuming that the distribution of income is initially optimal (such an assumption is clearly necessary for pure forms of capitalization). Next, we consider an initial situation and a proposed new project. Some form of capitalization occurs when there is a correlation between the resulting welfare change and the associated rental change. We will see that this correlation can take many forms.

All precise statements about internal capitalization require some boundary condition on community rents. This boundary condition will take different forms, depending on the structure of the town and the organization of taxation in the town. We will discuss two different types of communities in this regard. In the first type of community the boundaries will be considered potentially variable, with "outside" land always available to the community at an exogenously given opportunity cost; the boundary rent is thus exogenously given in such a community. The real prototype for such a community would be a town or metropolitan area located in the middle of farmland. We will refer to such communities as isolated.

In the second type of community, the boundaries are predetermined (presumably by legal arrangement). The effect of a project on

boundary rent will be determined by changes (if any) in the aggregate demand for land in the community. The real prototype here would be a subcommunity within a larger metropolitan area. We will refer to such communities as adjacent.

The local public goods model being proposed here seems most applicable to the case of isolated communities, since it is difficult to ignore the importance of direct spillovers among adjacent communities. However, it still seems useful to treat the adjacent case.

To demonstrate that there are separate forces generating external and internal capitalization, we will need to show that it is possible to have each without the other. We will begin with a discussion in which we assume that no external forces are present; that is, we assume that community policy will not lead to any net migration to or from the community. Within this context, we explore the conditions for internal capitalization. We derive alternative sets of assumptions under which there is no capitalization or full capitalization.

Then we will return to the issue of external forces. We will give these forces their best chance to work by assuming free mobility between communities. To isolate the effect of these forces, we will adopt a model of the community which is consistent with no internal capitalization. Within this context, we again exhibit conditions under which there is either no capitalization or full capitalization.

A. *Internal Capitalization*

Assuming away migration effects, there are essentially two ways in which project benefits can get translated into increased land values: through a change in the extensive margin or the intensive margin. If a project increases the aggregate demand for land, it may lead to an increase in the general level of rents (by operating on the extensive margin for land). Intuitively, this force could go either way. When taxes are raised (to pay for a new project), there is likely to be a substitution away from goods (including land). On the other hand, when public goods provision is increased, there is likely to be a complementary increase in the demand for land. Since the net effect on rents is clearly ambiguous, there is no reason to expect systematic capitalization from impact on the extensive margin. In fact, we will argue that the extensive margin is likely to be of no importance, especially in isolated communities.

However, the project may affect the differential value of locations and thereby influence the rent structure through the internal margin. We will show that this force may lead to systematic capitalization in relatively large communities, within which there is a wide range of choice regarding use of the public goods. Restricting this range of

choice can lead to any intermediate case from full capitalization to no capitalization.

B. *Sufficient Conditions for No Internal Capitalization*

The essential restriction which is needed to rule out internal capitalization is that there be no effective differential choice in the use of public goods within the community. A number of sets of assumptions will do for this purpose.

The easiest case to explain is one in which there is no differential choice within the community at all. By this, we mean that $g = \bar{g}$ and $\ell = \bar{\ell}$ for all agents in all zones; these conditions would hold if the plot size is institutionally given and the public good has the characteristics of national defense so that residents get the benefits regardless of their actions.

A resident of preference type a and income class i now chooses s (the only remaining choice variable) to $\max_s U^a[q,\bar{g},\bar{\ell},I^i - T^i - f(\bar{g},s) - r_s\bar{\ell}]$. Clearly, this problem reduces for every agent to one of minimizing costs: $\min_s [f(\bar{g},s) - r_s\bar{\ell}]$. Since the costs to be minimized are the same for all agents, the only possible equilibrium rent structure must be one which makes those costs independent of s. (At such a cost structure, everyone is obviously indifferent as to where they live.) Now, this independence condition must be true both before and after the project is initiated. It follows that the derivative of costs with respect to q must be independent of s. But with g and ℓ fixed, this means that $(d/dq)r_s$ must be independent of s; that is, if rents change at all they must change in a uniform way. However, rents at the boundary cannot change: There is no change in the aggregate demand for community land, so the boundary rent will not change in either isolated or adjacent communities. Hence there is no capitalization. Thus we can identify one case which formalizes the assertions of Hamilton (1975).

We can introduce further choice into the community and obtain the same results as long as we are willing to restrict the degree of diversity and the degree of complementarity of tastes. For example, suppose that tastes are separable and similar in that preferences of all agents can be represented in the form $U = \hat{U}^1(q,g) + \hat{U}^2(\ell) + I^i - T^i - f(g,s) - r_s\ell$, where i indexes incomes class, as before. Now, if it should turn out that the optimal choice of g is independent of s (so that there is no effective differential choice within the town), we can again argue that there will be no capitalization. Clearly agents still make the same choices regardless of income class. Thus, rents must still adjust so that everyone is indifferent as to where they live (V_s^i is independent of s). This condition must be true both before and after the project, so

$(d/dq)V^i_s$ must be independent of s. Differentiating and using the envelope theorem, this condition becomes that

$$\frac{d}{dq}\,\hat{U}(q,g_s) + \ell_s\,\frac{dr_s}{dq} \tag{2}$$

is independent of s. But since we assumed that g_s was independent of s, the first term of (2) is independent of s, and we are back to the statement that the rent structure must change uniformly if at all. In fact, it cannot change at all, because there is no change in the extensive demand for land (given separable preferences).[2]

C. *Sufficient Conditions for "Full" Internal Capitalization*

Internal capitalization can take a number of different forms, depending on specific assumptions made concerning the boundary conditions, the type of taxes imposed, and the nature of property ownership. We will delineate these cases as we go along. As will be seen, all of the propositions reported here rely exclusively on the intensive land margin to "enforce" capitalization. There may be some very special functional forms which will generate systematic capitalization from the extensive margin, but these would seem too special to be of much interest.

There are two principal assumptions which taken together guarantee that project benefits will be translated to land rents through the intensive margin. The first is that all the benefits of the project must be "intramarginal" in that boundary residents are marginally unaffected. We would expect this condition to hold if marginal residents choose $g = 0$ (so that there is a complete revealed range of choice within the community). But the condition they may hold more generally; for example, in the case of a museum or park, "boundary" residents may make so few trips that, at their current level of activity, they could make no better use of a larger facility. We formalize this condition as follows. Let σ stand for some boundary region. Then "marginal indifference" will mean that

$$\frac{\partial}{\partial q}\,U[q,g^i_\sigma,\ell^i_\sigma,I^i - T^i - f(g,\sigma) - r_\sigma \ell^i_\sigma] = 0. \tag{3}$$

The second necessary assumption is that residents do not tend to sort themselves out within the town according to their relative prefer-

[2] This case is closest in spirit to the example of no capitalization given in Polinsky and Shavell (1976). It also corresponds to Hamilton's (1975) model of Tiebout communities in which there is no internal spatial structure. As we show below, Hamilton's conclusion that there would be no capitalization in Tiebout communities depends critically on the assumption of no internal spatial structure.

ence for the public good. Since systematic differences will generally lead to sorting, this means that the community must be reasonably homogeneous in attitudes toward public goods. We will see later that, when sorting occurs, it will tend to mitigate capitalization.

The required homogeneity will result whenever agents are additively separable in their preferences on any component of consumption concerning which they systematically differ. Here we will treat the case where individuals differ in incomes and have preferences which are additively separable in the associated numeraire consumption good.

Consequently, we now assume that a resident (i) of the community has preferences which can be represented in the form $U^i = \hat{U}(q,g,\ell) + I^i - T^i - f(g,s) - r_s\ell$. Now, it is obvious that residents will not sort themselves out by income class. Conditional on living at s, all residents would make the same choices of g and ℓ ($g_s^i = g_s$, $\ell_s^i = \ell_s$, all s), so at any specified rent structure, they would all agree on the best location; hence, as before, the rent structure must adjust until all locations are equally desirable for all residents.

We now evaluate the first-order welfare effect of a new project:

$$dW = \sum_i \omega_i dV^i. \tag{4}$$

Since we are ignoring distributional factors, we must assume that whatever income differences prevail are "optimal," implying that the welfare weights (ω_i) are equal (otherwise, pure transfers could be made in such a way as to improve net welfare). We normalize units by setting the common weight equal to one. In evaluating (4) we treat everyone as if they were living at the boundary region σ and take into account all potential effects which q could have on the parameters faced by consumers.

Performing the differentiation, applying the envelope theorem where appropriate, and aggregating where possible, we have

$$\frac{dW}{dq} = N \frac{\partial}{\partial q} \hat{U}(q,g_\sigma,\ell_\sigma) - N\ell_\sigma \frac{dr_\sigma}{dq} + \frac{dY}{dq} + \frac{d\Pi}{dq} + \frac{dR}{dq} - \frac{dT}{dq},$$

where N stands for the total population of the community and the income variables without superscripts stand for aggregates over all residents. Now, marginal indifference (3) means that the first term is zero, and dY/dq must be zero as well, since Y cannot change under the assumption that nonland private goods prices are fixed. Furthermore, government budget balance implies that $dT/dq = d\Gamma/dq$. Making these substitutions, we have, finally,

$$\frac{dW}{dq} = \frac{dR}{dq} - \frac{d\Gamma}{dq} + \frac{d\Pi}{dq} - N\ell_\sigma \frac{dr_\sigma}{dq}. \tag{5}$$

The formula (5) can be thought of as a generic capitalization result from which a variety of specific relationships can be derived.

1. Local Ownership

Let us first look at the case where local firms (as well as local land) are all owned by local residents. In that case Π should be thought of as total profits generated in the region. Assuming that local firms behave competitively, we know that $d\Pi$ can be written in the form $d\Pi = X \cdot \delta p$, where X is the vector of net outputs and p the private goods price vector. But the only prices which change are land prices, so we have $d\Pi/dq = -(dR_f/dq)$, where R_f is the value of land used by firms. Substituting this relationship into (5), we have

$$\frac{dW}{dq} = \frac{dR_r}{dq} - \frac{d\Gamma}{dq} - N\ell_\sigma \frac{dr_\sigma}{dq}, \tag{6}$$

where R_r is the value of residential land.

Alternatively, we could take the view that firms exhibit constant returns to scale and are constantly in long-run equilibrium. In this case, the above analysis is invalid since any infinitesimal change in prices will lead to a discrete jump in the firm's decisions; in particular, if rents go up, ceteris paribus, all firms go out of business. The correct analysis would then set $d\Pi = 0$ in (5) and conclude (as in most of the previous examples in the literature) that capitalization involves all land value. Upon reflection, it seems that the long run is too long to be relevant here. Firms are not free to leave when rents rise, and the increasing rent payments by firms should be counted against profits. Hence we will stand by the formulation (6).

The final form capitalization takes depends on the boundary conditions. In the isolated case, it seems reasonable to suppose that $dr_\sigma/dq = 0$ even though the boundary of the town may shift. Conceptually, we can think of there being many boundary regions which are shared by residents and farmers. As long as some of these regions are still shared after the project is initiated, boundary rent cannot change. Thus for the case of isolated communities, we assert $dW/dq = (dR_r/dq) - (d\Gamma/dq)$. Verbally, the gross benefits of the project are capitalized into residential land values. Net benefits (dW/dq) are gross benefits minus costs $(d\Gamma/dq)$.

A more general statement can be made which will apply to adjacent communities as well. If the aggregate demand for land increases so that $dr_\sigma/dq > 0$, then the value of residential land overcapitalizes gross benefits ($[dR_r/dq] - N\ell_\sigma[dr_\sigma/dq]$ exactly capitalizes gross benefits) while if $dr_\sigma/dq < 0$, then the value of residential land undercapitalizes gross benefits.

2. National Ownership

The majority of firms are not locally owned. Indeed, most firms are national in the scope of their operations, and local ownership has no meaning for such firms. Suppose we take the position that all ownership is in national firms. What happens to the analysis? Actually, not very much. Indeed, the formula (6) is still correct for society as a whole, but now some of the costs of the project are paid by "foreigners" who own shares in the firms which are operating locally. Indeed, if we take the position that each community is small relative to the country, then it is a good approximation to assume that local branches of national firms are owned entirely by outsiders. In that case, it follows that gross benefits to residents of the community are capitalized into total land value, while gross benefits to society as a whole are capitalized into residential land values. The difference represents an externality. We will discuss other examples of such externalities later.

3. Property Taxes

Up until now we have dealt with only the case in which taxes were treated as lump-sum taxes by consumers. Since the major locally imposed tax is a property tax, it is important to see how the analysis needs to be modified in that case. If an ad valorem tax at rate t is imposed, then an agent using an amount of land ℓ in zone s will pay taxes $tr_s\ell$, where r_s must now be thought of as the rent net-of-tax payments.

Now, the problem for a typical individual can be stated as $\max_{g,\ell,s}$ $[\hat{U}(q,g,\ell) + I^i - f(g,s) - r_s(1+t)\ell]$. Clearly, it is still true that all agents will make the same choices, and, indeed, the method of analysis is exactly as before. Performing the calculations (and taking the point of view of society in order to avoid any ambiguity concerning the ownership of firms), we derive

$$\frac{dW}{dq} = \frac{dR_r}{dq} - Nr_\sigma \ell_\sigma \frac{dt}{dq} - N\ell_\sigma(1+t)\frac{dr_\sigma}{dq}, \tag{7}$$

where R_r now stands for total residential rent net of taxes.

The exact form of capitalization again depends on boundary conditions. And now there is some ambiguity concerning these even in the case of isolated communities. The issue revolves around whether farmers in the boundary regions do or do not pay the taxes. If they do pay the taxes, then the free boundary condition is as before. However, if they do not pay the taxes (but would have to if they acquired

property "within" the community), then the appropriate condition is $r_\sigma(1 + t) = \bar{r}$, where $\bar{r}$ is the opportunity cost of land to a farmer. We will refer to this arrangement as *agricultural zoning* and to the other case as *blind zoning.*

The case of agricultural zoning is easiest to analyze. Differentiating the boundary condition with respect to q, we have $(1 + t)(dr_\sigma/dq) + r_\sigma(dt/dq) = 0$. Therefore, referring back to (7),

$$\frac{dW}{dq} = \frac{dR_r}{dq}. \tag{8}$$

Here net project benefits are capitalized into net-of-tax residential property values, while gross project benefits are capitalized into before-tax property values.

For the case of blind zoning, the term involving dr_σ/dq vanishes in (7), and we are left to evaluate $Nr_\sigma\ell_\sigma(dt/dq)$. Suppose that only residential land is taxed, so that the government budget balance condition is $tR_r = \Gamma$. Differentiating this equation with respect to q and substituting for dt/dq in (7) yields

$$\frac{dW}{dq} = \frac{dR_r}{dq} - \alpha\left(\frac{d\Gamma}{dq} - t\,\frac{dR_r}{dq}\right) = (1 + \alpha t)\,\frac{dR_r}{dq} - \alpha\,\frac{d\Gamma}{dq}, \tag{9}$$

where $\alpha = N\bar{r}\ell_\sigma/R_r$. The net benefit is now a weighted difference between increases in after-tax residential property values and increases in cost. Note that the smaller is α, the closer we approximate the results for the previous case. The α stands for the ratio of the value of land residents would use if they all lived at the boundary to the actual value of residential land.

More generally, however, (9) tells us that increases in residential property value overcapitalize net benefits by an amount related positively to the gap between extra costs and the extra revenue that would be generated without changing tax rates. Clearly, this result is intermediate between pure gross capitalization and pure net capitalization.

D. *Preference Differences within the Community*

Once systematic preference differences are introduced into the communities, the exact capitalization results tend to break down. Here, we demonstrate this and indicate the type of modification which is required. For this purpose, let us consider the simplest extension possible, one with two different types of residents in the town (indexed a and b). Let there be N_a (N_b) a-type (b-type) residents. It is convenient to assume that all a-type residents have the same preferences and income (although different incomes could be incorporated as before).

Also, we find it necessary to suppose that the plot size within any given zone must be uniform, and we may as well assume that this plot size is predetermined.

Now we let S^a stand for the set of zones occupied by a-type residents and S^b for the corresponding set of zones for b-type residents. Obviously if σ is included in both sets, the analysis is exactly as in previous sections; indeed, all previous theorems go through with preference differences as long as those differences do not lead to systematic sorting. Therefore, without loss of generality, we can assume that the zoning arrangements are as in the Venn diagram shown in figure 1.

Since the zoning is discrete, there is little loss of generality in assuming that at least one zone is occupied by both types. We let τ stand for such a zone.

In what follows, it is convenient to use the shorthand notation $F_s^x(q) = \hat{U}^x(q,g_s^x,\ell_s) - f(g_s^x,s)$, all s, $x = a,b$. Thus, V_s^a can be written in the form $V_s^a \equiv F_s^a(q) - r_s\ell_s + I^a$. Next we evaluate welfare as before, except that now we cannot evaluate everyone's welfare at σ; instead, we evaluate a-type welfare at σ and b-type welfare at τ:

$$\begin{aligned} W &= N_a V_\sigma^a + N_b V_\tau^b, \\ &= N_a F_\sigma^a + N_b F_\tau^b + I - N_a r_\sigma \ell_\sigma - N_b r_\tau \ell_\tau. \end{aligned} \tag{10}$$

Finally, we make use of the fact that a-type agents are indifferent between locations τ and σ. This indifference implies that

$$r_\tau \ell_\tau = r_\sigma \ell_\sigma + F_\tau^a(q) - F_\sigma^a(q). \tag{11}$$

Substituting (11) into (10) yields $W = NF_\sigma^a(q) + N_b[F_\tau^b(q) - F_\tau^a(q)] + I - Nr_\sigma\ell_\sigma$, and differentiating with respect to q, we obtain:

$$\begin{aligned} \frac{dW}{dq} &= N\frac{d\hat{U}^a}{dq}(q,g_\sigma^a,\ell_\sigma) - N\ell_\sigma\frac{dr_\sigma}{dq} + \frac{d\Pi}{dq} + \frac{dR}{dq} - \frac{dT}{dq} \\ &+ N_b\left[\frac{d\hat{U}^b}{dq}(q,g_\tau^b,\ell_\tau) - \frac{d\hat{U}^a}{dq}(q,g_\tau^a,\ell_\tau)\right]. \end{aligned} \tag{12}$$

Examining (12), we see that the terms are just as before except for the intramarginal benefit term:

$$N_b\left[\frac{d\hat{U}^a}{dq}(q,g_\tau^b,\ell_\tau) - \frac{d\hat{U}^a}{dq}(q,g_\tau^a,\ell_\tau)\right]. \tag{13}$$

We will argue that there is a strong presumption that this term is positive. To begin with, the condition $g_\tau^b > g_\tau^a$ is a stability condition for the town structure we have specified. It guarantees that there is a differential benefit to an a-type agent relative to a b-type agent as one moves toward the boundary of the town (b-type agents gain relatively more from proximity to the center since they make more trips).

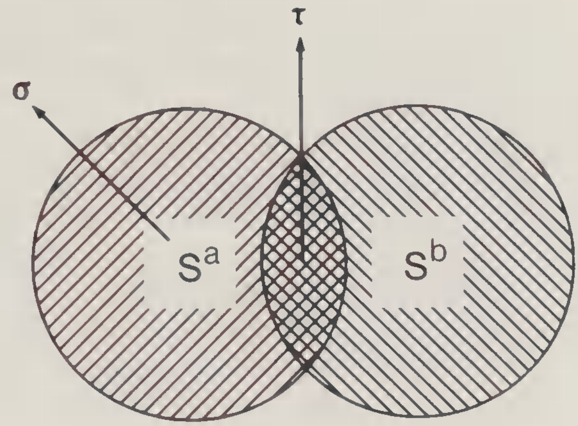


FIG. 1

Therefore, as long as the marginal benefit from more public goods increases with the number of trips, the intramarginal term (13) must be positive.

Thus, when systematic preference differences are introduced into any particular variant of the model studied above, land rents will tend to undercapitalize the associated benefit. Operationally, this means that if land rents were used as a decision criterion, we might end up rejecting projects which ought to be accepted.

E. *Sufficient Conditions for No External Capitalization*

Having explored the conditions for internal capitalization in a community that was insulated from immigration forces, we turn now to a study of those external forces. We will show first that any systematic differences between towns can serve to neutralize the external forces at least with respect to their effect on small projects. There is an important caveat here, however. What we show is that communities could arrange things so as to insulate themselves; we cannot show that they have an incentive to follow such policies. Indeed, we have shown in some related work that communities do not have a myopic incentive to insulate in many instances.[3] We will ignore this issue here.

We employ the simplest model of community here, in which there are no choices with regard to public goods use. The reader might usefully recall that this is one of the cases in which there is no internal capitalization. All agents within the town are assumed alike. To make the point about potential insulation in the strongest possible way, we will also assume that preferences of all individuals in all towns are alike and that towns differ only in income levels; that difference alone

[3] See Starrett 1977.

is enough to insulate towns. Actually, this case may not be so unrealistic, since sorting by income level is the clearest observed Tiebout differentiation among communities.

Referring back to the first model in Section II*A*, we can specify the characteristics of community k completely by a level of public goods (q^k), an income level of residents ($\bar{I}^k$),[4] and a cost level ($h^k = T^k + f^k[\bar{g}^k,s] + r^k_s\hat{\ell}_s$) which must, at equilibrium, be independent of location (s). It is convenient to assume that the $\bar{g}$ and $\hat{\ell}$ levels are the same everywhere so that they can be deleted from the analysis, although this is not really necessary. Note that, in the model as now constituted, there is no economic distinction between property taxation and lump-sum taxation. Now, the welfare (U^{kk}) of an individual with income $\bar{I}^k$ living in his own community can be expressed as $U^{kk} \equiv U(q^k,\bar{I}^k - h^k)$.

We now ask the following question. Suppose that a set of q's, I's, and h's is specified. Will anyone have an incentive to move immediately, and if not would anyone see such an incentive if one of the q's were changed marginally? We will show that an appropriate initial choice will make the answer to both these questions "no" as long as income effects are not negligible. Suppose a k agent considers moving to j. We assume that, if he does this, he must pay the taxes (and other costs) appropriate to j. (If, instead, he expects to pay the taxes appropriate to k, as he surely would if the tax were a property tax, the results are still the same, but we would have to incorporate the taxes in the I term rather than the h term.) On the other hand, an agent's income is determined by his place of initial ownership. This must be true of property income (as long as there are no unanticipated capital gains), and labor income is the same everywhere, by assumption. Hence, he would expect to get utility $U^{kj} \equiv U(q^j,I^k - h^j)$.

The move is not desirable, and would not be desirable for any sufficiently small variation in q^j, as long as $U^{kk} > U^{kj}$. Similarly, a move from j to k is not desirable as long as $U^{jk} \equiv U(q^k,\bar{I}^j - h^k) < U(q^j,\bar{I}^j - h^j) \equiv U^{jj}$. Are these two conditions consistent? Let us order j and k in such a way that $q^k > q^j$. Then stability certainly requires $h^k > h^j$. To see what else is required, it is convenient to change variables slightly. If we define $\Delta q = q^j - q^k$, $z^k = \bar{I}^k - h^k$, $z^j = \bar{I}^j - h^k$, and $\Delta z = h^k - h^j$, then the two stability conditions may be written as $U(q^k,z^k) > U(q^k + \Delta q,z^k + \Delta z)$ and $U(q^k,z^j) < U(q^k + \Delta q,z^j + \Delta z)$, with $\Delta q < 0$ and $\Delta z > 0$. These conditions are consistent as long as $\bar{I}^k > \bar{I}^j$, as can be seen in figure 2. Of course, we have assumed in the construction that the income effect is normal and significant, so that increases in income increase the marginal rate of substitution of z for q. Clearly, community k could

[4] The $\bar{I}^k$ is clearly the average income in the community.

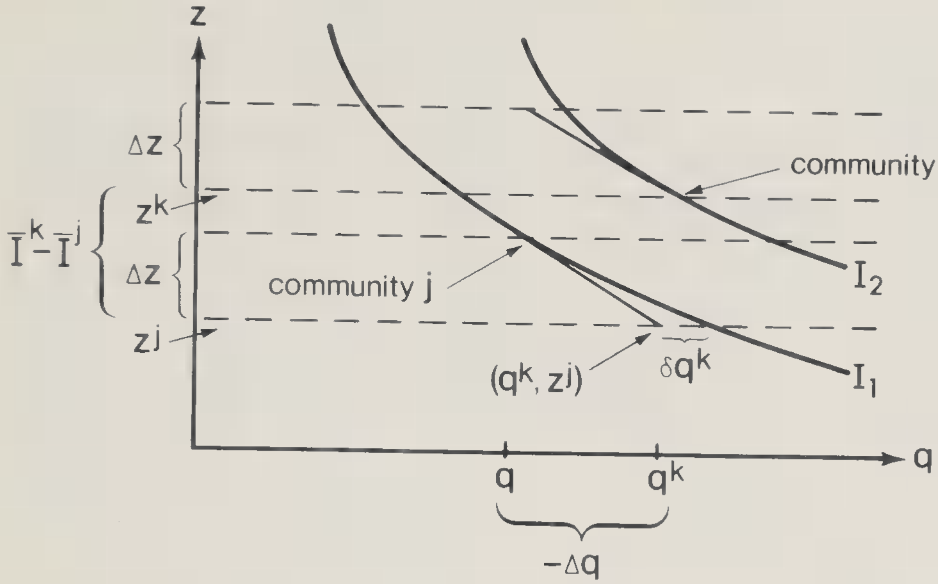


FIG. 2

engage in any project of size less than or equal to δq^k without inducing entry, while a similar statement holds for community j.

Naturally, if projects undertaken are too large, entry may be induced. However, we assert that any systematic differences between towns (direct preference differences would have done as well) can imply no external capitalization of marginal projects. This same threshold effect would result if the segregation of communities were caused by a fixed cost of moving (rather than by preference differences).[5]

It is interesting to note that if we now reintroduce the possibility of internal capitalization, its presence may be enough to neutralize external forces, even if there are no intrinsic differences at all among people. That is, if $q^k > q^j$ implies $R^k > R^j$, we will automatically have $\bar{I}^k > \bar{I}^j$, assuming that there are no other income differences.

We chose income differences to illustrate the possibility of insulation because they were the most innocuous-sounding differences we could think of. Obviously, preference differences are more likely to serve this purpose.[6] Indeed, some might argue that income effects are unimportant in the context discussed here. Suppose that we reexamine our problem in the absence of income effects. Interestingly, the whole character of the analysis changes.

[5] Of course, if there are no costs of moving and a continuum of preference types, any project will induce some entry and, presumably, some capitalization.

[6] Since the Tiebout model is characterized by preference difference among communities, there will be no external capitalization of small projects in Tiebout communities that are "sufficiently" differentiated.

F. Sufficient Conditions for Full External Capitalization

Without income effects or preference differences between communities it is impossible to insulate, and the resulting migration forces will generate external capitalization. To show this, let us reintroduce preference differences within communities while at the same time eliminating any income effects or preference differences between communities. An a-type resident of income class i now has preference of the form $\hat{V}^{ai} = \hat{U}^{a}(q) + I^{i} - h$. At this point, it is convenient to define a fictitious "average" resident of community k. This average resident clearly would have preference of the form $\bar{V}^{k} = \hat{\hat{U}}^{k}(q^{k}) + \bar{I}^{k} - h^{k}$, where $\bar{I}^{k}$ is the average income level in k (as before) and $\hat{\hat{U}}^{k}$ is the average felicity function in k. Our assumption of no preference differences between communities will be taken to mean that the average felicity functions are the same in all communities, that is, $\hat{\hat{U}}^{k}(q) = \hat{\hat{U}}(q)$, all k.

Now let us reconsider the stability conditions for equilibrium. Clearly, a necessary condition for stability is that our fictitious average resident does not want to move in either direction between any pair of communities. The corresponding pair of inequalities for communities k and j are

$$\hat{\hat{U}}(q^{k}) + \bar{I}^{k} - h^{k} \geq \hat{\hat{U}}(q^{j}) + \bar{I}^{k} - h^{j} \tag{14}$$

and

$$\hat{\hat{U}}(q^{k}) + \bar{I}^{j} - h^{k} \leq \hat{\hat{U}}(q^{j}) + \bar{I}^{j} - h^{j}. \tag{15}$$

Clearly strict inequality in either direction is impossible, and both relations must hold as equalities in any equilibrium situation; the average resident must be indifferent at the margin as to where to live. It follows that any welfare-improving project anywhere must induce entry.

We can now use equations (14) and (15) to generate a simple external capitalization result. These equations must hold both before and after a project (in some particular town z), so we can differentiate them with respect to q^{z}. In doing so we must allow for the possibility that a change in q^{z} could affect I^{j} and h^{j} ($j \neq z$). We have already seen one way in which this could happen through external ownership of local firms; now that migration is certain, there are the so-called fiscal externalities as well; when one community attracts residents from other communities, the lost tax revenue is a real external cost to the "losing" communities. For a rigorous demonstration of this fact, the reader is referred to Starrett (1980).

Allowing for potential changes in all variables, we can write

$$dV^{z} \equiv d\hat{\hat{U}}(q^{z}) + d\bar{I}^{z} - dh^{z} = d\bar{I}^{z} - dh^{j} \tag{16}$$

and

$$dV^j = d\hat{\bar{U}}(q^z) + d\bar{I}^j - dh^z = d\bar{I}^j - dh^j, \tag{17}$$

where dV^i is shorthand notation for the change in average welfare of community i.

Let us define $dW^\#$ to be the total external effect; that is,

$$dW^\# = \sum_j N^j dV^j = dI^\# - \sum_{j\neq s} N^j dh^j.^7$$

And $dW^z = N^z dV^z$ will stand for the welfare change in community z. Clearly, $dW = dW^\# + dW^z$. Simple substitutions yield the following two equations for the change in community z welfare and the change in social welfare:

$$dW^z = \frac{N^z}{N^\#} dW^\# - \frac{N^z}{N^\#}(dR^\# + d\Pi^\#) + dR^z + d\Pi^z$$

and

$$dW = \frac{N}{N^\#} dW^\# - \frac{N^z}{N^\#}(dR^\# + d\Pi^\#) + dR^z + d\Pi^z.$$

Finally, we must distinguish cases again, according to the nature of firm ownership.

1. Neutral Ownership

Suppose that each community owns its share of national profits. Then $N^z\Pi^\# - N^\#\Pi^z = 0$, and we have

$$dW^z = \frac{N^z}{N^\#} dW^\# + dR^z - \frac{N^z}{N^\#} dR^\# \tag{18}$$

and

$$dW = \frac{N}{N^\#} dW^\# + dR^z - \frac{N}{N^\#} dR^\#. \tag{19}$$

Clearly, total rent change in z is a good approximation to net welfare benefits in z as long as community z is small relative to the nation (so that $N^z/N^\#$ is small). However, net social benefits will be misrepresented by rent increases to the extent that the external effect is significant.

How important is this misrepresentation? It turns out it may be

[7] The I^i stands for total income in community i; the symbol # means the associated variable is aggregated over all communities except z. Recall that we are assuming that all local public goods projects are paid for locally. Naturally we could correct for the externalities if we could impose the optimal revenue-sharing scheme.

quite important even if $N^z/N^{\#}$ is small. We have studied the measurement of $dW^{\#}$ elsewhere (see Starrett 1980) and merely report the main results here.

When community z engages in a project, that project is certain to induce some entry. Other communities suffer real welfare losses which are equal to lost tax revenue (from the migrators) minus any attendant improvement in congestion "costs." Unless communities have already seriously overexpanded, the net benefit to outsiders is negative ($dW^{\#} < 0$). Hence even changes in residential land values will overstate net social benefit, quite generally.

Of course, if community z is not small as a fraction of the country, the theory takes a somewhat more complicated general form. Aside from the external terms, net benefits are capitalized into the difference between the change in land value in z and the "average" change in land values elsewhere. This formulation is reminiscent of that presented in Strotz (1968) for two regions.

There is one other major difference from the internal capitalization results: net benefits are externally capitalized, while gross benefits were internally capitalized. The economic reason for this discrepancy is not difficult to find. The marginal potential resident pays local taxes in the internal model (he is the boundary resident), while the marginal potential resident does not pay local taxes in the external model.

2. Local Ownership

If each community owns its own firms, then $d\Pi^{\#} = -dR_s^{\#}$, and we have

$$dW^z = \frac{N^z}{N^{\#}} dW^{\#} + dR_r^z - \frac{N^z}{N^{\#}} dR_r^{\#} \tag{20}$$

and

$$dW = \frac{N}{N^{\#}} dW^{\#} + dR_r^z - \frac{N^z}{N^{\#}} dR_r^{\#}. \tag{21}$$

The form of the results is the same except that capitalization is into residential rather than total land value. The reader might think at first that there is a contradiction between equations (19) and (21). After all, it cannot matter from a social point of view where the profits are owned, so why do these equations look different? The answer is that any increased cost of local industrial land will be reflected in the term $dW^{\#}$ in equation (19) but not in (21). Thus, even in the case of national ownership, only increases in residential land values represent real social benefit, once adverse land value externalities are canceled out.

III. Summary

We have shown in the previous sections that there is considerable diversity in the form and extent of capitalization, depending on the nature of the assumptions made. However, some fairly general principles emerge, and we summarize these in this section.

First, all of the models studied lead to the conclusion that, from the point of view of society as a whole, it is residential property values that capitalize project benefits (rather than total land value). The theories differ on the appropriate land base only from the perspective of the project-generating community; if lost profits from increased land values are exported, then increases in total land value measure benefits to the community, but, if not, then it is residential land values.

The internal and external capitalization models generally disagree on whether gross or net benefits are capitalized. However, it is worth pointing out one important case in which the two theories agree on this issue. This is the case of property taxation with agricultural zoning, for which both theories suggest that increases in net-of-tax land values capitalize net welfare benefits (and, naturally, increases in gross-of-tax land values capitalize gross benefits).

The main differences between internal and external capitalization results can be traced to the boundary conditions imposed on a representative community. Notice that such boundary conditions play no role at all in external capitalization; indeed, the external results hold regardless of what happens to rents or benefits at the boundary of the communities.

However, internal capitalization results vary considerably, depending on the assumptions governing behavior of land rent and the impact of incremental projects, at the boundary. In the extreme case in which the impact is nil and the (net-of-tax) rent goes down by the rate of taxation, internal capitalization agrees with external capitalization. But if boundary rent is unchanged (as it will be under many plausible conditions) the gross project benefits are internally capitalized. And if marginal projects actually make boundary residents worse off (because of a dominating congestion effect) then land values will overcapitalize even gross benefits.

We should recognize that, in all cases considered, land values will fail to capture some of the external benefits (or, more likely, costs) of a project. When the only external effect is through nationally owned profits, as was the case in the pure internal capitalization model, we could correct for the externality by using residential land value rather than total land value. However, in the free-trade, free-migration model of external capitalization, the external costs of a project are

generally much larger, and we conclude that increases in residential land value will overstate true net social benefit by an amount closely related to the size of the fiscal externality.

We close with some remarks on the basic assumptions underlying capitalization results. The two types of capitalization rest on different, and to some extent complementary, assumptions. External capitalization requires that the distribution of agents by economic characteristics be the same in one community as in all others, while internal capitalization requires that agents within a particular community must have similar economic characteristics (though these characteristics may differ from community to community). Thus, the internal capitalization results are most applicable in a "pure Tiebout" world where each community specializes and attracts a single type of agent. The external capitalization model is most applicable in the opposite-extreme world of "pure scrambling" in which all communities look alike in terms of the characteristics of agents. Naturally, there are many intermediate cases between these extremes (including the real world, presumably!). To the extent that neither extreme is a good approximation, capitalization will not hold in either form, or will hold only partially.

References

Freeman, A. Myrick, III. "Air Pollution and Property Value: A Methodological Comment." *Rev. Econ. and Statis.* 53 (November 1971): 415–16.

Hamilton, Bruce W. "Property Taxes and the Tiebout Hypothesis: Some Empirical Evidence." In *Fiscal Zoning and Land Use Controls: The Economic Issues*, edited by Edwin S. Mills and Wallace E. Oates. Lexington, Mass.: Lexington, 1975.

Lind, Robert C. "Spatial Equilibrium, the Theory of Rents, and the Measurement of Benefits from Public Programs." *Q.J.E.* 87 (May 1973): 188–207.

Oates, Wallace E. "The Effects of Property Taxes and Local Public Spending on Property Values: An Empirical Study of Tax Capitalization and the Tiebout Hypothesis." *J.P.E.* 77, no. 6 (November/December 1969): 957–71.

Pines, David, and Weiss, Yoram. "Land Improvement Projects and Land Values." *J. Urban Econ.* 3 (January 1976): 1–13.

Polinsky, A. Mitchell, and Rubinfeld, Daniel L. "The Long Run Effects of a Residential Property Tax and Local Public Services." *J. Urban Econ.* 5 (April 1978): 241–62.

Polinsky, A. Mitchell, and Shavell, Steven. "Amenities and Property Values in a Model of an Urban Area." *J. Public Econ.* 5 (January/February 1976): 119–29.

Starrett, David A. "On the Capitalization Hypothesis in Closed Communities." Technical Report no. 241. Inst. Math. Studies Soc. Sci., Stanford Univ., 1977.

———. "On the Capitalization Hypothesis for Local Public Finance." Unpublished manuscript, Stanford Univ., June 1978.

———. "On the Method of Taxation and the Provision of Local Public Goods." *A.E.R.* 70 (June 1980): 380–92.

Strotz, Robert H. "The Use of Land Rent Charges to Measure the Welfare Benefits of Land Improvement." In *The New Economics of Regulated Industries: Rate-Making in a Dynamic Economy*, edited by Joseph E. Haring. Econ. Res. Center, Occidental College, Los Angeles, 1968.

Tiebout, Charles M. "A Pure Theory of Local Expenditures." *J.P.E.* 64, no. 5 (October 1956): 416–24.

Money and the Dispersion of Relative Prices

Zvi Hercowitz

University of Rochester and National Bureau of Economic Research

A price dispersion equation is tested with data from the German hyperinflation. The equation is derived from a version of Lucas's and Barro's partial information-localized market models. In this extension, different excess demand elasticities across commodities imply a testable dispersion equation, in which the explanatory variable is the magnitude of the unperceived money growth. In order to test this hypothesis a price dispersion series is constructed, and a measure of the unperceived part of money growth is estimated. The model receives support from the empirical analysis, although it is evident that unincluded variables have important effects on price dispersion.

The existence of a positive correlation between absolute price level variability and the dispersion of relative prices was observed by, for example, Mills (1927), Graham (1930), and recently by Vining and Elwertowski (1976). In his study of U.S. price behavior during the period 1920–26, Mills (1927) says: "We have not however exhausted the possibility of discovering a relationship between price level and dispersion. It may be that dispersion depends upon the violence of the price change, regardless of direction" (p. 284). Graham (1930) finds in the post–World War I German hyperinflation an additional dynamic element of price behavior: "It is clear that with the initiation of an upward movement in general prices a series of lags in individual

This paper is part of my doctoral dissertation at the University of Rochester. I wish to express my gratitude to Robert Barro for his guidance and helpful suggestions. I have also benefited from discussions with Bill Bomberger, Rudiger Dornbusch, Chip Miller, Mark Rush, Hal White, and from insightful comments by a referee of this journal. Financial support from the National Science Foundation is gratefully acknowledged.

[*Journal of Political Economy*, 1981, vol. 89, no. 2]

prices developed, that these lags tended quickly to disappear when stability of general prices was reached on a new level, or when general prices fell, but that they were nevertheless progressively eliminated even though the general price level continued to rise" (p. 175). This observation suggests that unexpected events may have an important role in the determination of price dispersion. Individual prices disperse at the beginning of an upward swing in the price level when the acceleration is presumably unexpected. As inflation continues the element of surprise wanes, and prices tend to converge.

The studies cited above failed to offer an economic rationale for the observed statistical correlation. Recently, a theoretical explanation of the relationship between price level variance and relative price dispersion was offered by Barro (1976). Using a localized markets framework of the type described by Phelps (1970) and employed by Lucas (1973), Barro links the dispersion of relative prices to the variance of the money supply. The key elements of this model are, on the one hand, individuals possessing incomplete current information and, on the other, demand and supply in each market reacting to relative prices as they are locally perceived. Thus, agents are confronted with the problem of determining whether locally observed price movements are caused by general inflation or by shifts in relative excess demand. The larger the variance of the money supply, the more likely are agents to attribute local price movements to general inflation rather than to relative shifts. Accordingly, as the money variance rises, local price changes induce smaller supply-and-demand responses—that is, excess demand becomes less elastic. Consequently, stochastic shifts to local excess demand produce larger changes in individual prices, so that the dispersion of prices across markets tends to increase with the variance of money. In this specification of the model, in which all markets have the same structure, dispersion is unrelated to the magnitude of realized money shocks.

This paper modifies Barro's framework by interpreting each location to be the market of a specific commodity, characterized by a particular excess demand elasticity. Because elasticities vary across markets, aggregate shocks affect each commodity price differently. Therefore, in this modified setup price dispersion is positively related to the magnitude of these shocks.

The model also predicts that systematic or perceived money growth is neutral with respect to price relationships. Accordingly, a money shock in this model is defined to be the component of money growth that is currently unobservable and cannot be inferred from currently available information. Whereas the quotation from Graham suggested that sudden—presumably unexpected—shifts in money growth cause dispersion, in this model unexpected monetary expan-

sion disperses prices only if it is, at least partially, currently unperceived.

The main task of this study is to evaluate this hypothesis with data from the German hyperinflation, a period of predominantly monetary disturbances. The period considered runs from January 1921 to July 1923. The vertiginous monetary expansion initiated in August 1923 differentiates the last phase of the hyperinflation, and, thus, it is not included in the sample.

The theoretical framework, presented in Section I, neglects some important facets of the hyperinflation. Specifically, it ignores the foreign exchange market and the sustained divergence between the internal and external values of the mark,[1] which obviously are related to relative prices. Also ignored in the main text are changes in the velocity of monetary circulation. However, in a brief discussion some general conditions are given under which price dispersion is neutral with respect to velocity changes. Finally, the only aggregate exogenous disturbances assumed to be affecting the economy are periodic infusions of new money made by the government. Real aggregate shocks are ignored. Empirically, they are probably of relatively minor importance and can be considered to be part of the error term in the estimated equations.

The testing of the dispersion equation requires two important preliminary steps. First, a price dispersion series is computed in Section II using an interesting set of data. It consists of monthly averages of 68 commodity prices, ranging from foods to metals, for the period of January 1921 through July 1923 (31 months). Data were unavailable for the months prior to January 1921. Next, a money growth equation is set up and estimated in Section III. To do so, both an information set available to agents economy-wide and a functional form relating this set to money creation are postulated and discussed. The explained part of money growth in the estimated equation is taken as a measure of the perceived rate of monetary expansion. Correspondingly, the unexplained part is interpreted as the money growth rate that could not be perceived from the assumed information set. These figures, in conjunction with the price dispersion series, are used in Section IV to test the price dispersion equation.

Parks (1978) has also tested a model of price dispersion using pre– and post–World War II U.S. data. In Parks's model dispersion is explained by changes in real income and the unexpected part of inflation, as measured by the innovation in the inflation rate. In this specification expected inflation and changes in real income are treated as exogenous variables. He finds a strong positive correlation

[1] See, e.g., Bresciani-Turroni 1937.

between unexpected inflation and price dispersion. His tests also suggest a separate but smaller effect of the actual inflation rate.

The present paper estimates an equation that relates price change dispersion to the exogenous shocks affecting the economy, in this case unperceived monetary injections. An additional monetary variable that is theoretically relevant for price dispersion is the variance of money shocks. An estimate for this variance is obtained from the money growth analysis and is included in the estimation. The model receives significant support from the empirical analysis. In particular, the variable measuring unperceived money growth has substantial explanatory power for price dispersion. The results also make clear that unincluded variables have important effects on price dispersion. Some of these are briefly considered in Section V.

I. The Model

The economy consists of an arbitrarily large number of physically separated markets indexed by z. In each location a specific commodity is produced and traded. At each date t the agents, assumed to be risk neutral, exchange money only for the commodity being traded in the market in which they are currently located. At date $t + 1$, agents change location at random, and the process is repeated. Consider now the information set available to the agents. It contains not only lagged values of all relevant variables, but also current information which is limited to the local market price $P_t(z)$, and some economy-wide shared knowledge about current variables related to money creation. Actual money growth, however, includes a random term which is assumed unknown.

The supply and demand for commodity z assume the log-linear forms:

$$y_t^s(z) = \alpha^s(z)[P_t(z) - EP_t] + \epsilon_t^s(z), \tag{1}$$

$$y_t^d(z) = -\alpha^d[P_t(z) - EP_t] + (M_t - EP_t) + \epsilon_t^d(z), \alpha^s(z) > 0, \alpha^d > 0. \tag{2}$$

The operator E is the mathematical expectation taken conditional on the information available in market z at time t. For each commodity z, $P_t(z) - EP_t$ is the locally perceived relative price. The expressions $\epsilon_t^s(z)$ and $\epsilon_t^d(z)$ represent relative shifts to supply and demand, respectively. The excess demand shift, $\epsilon_t(z) \equiv \epsilon_t^d(z) - \epsilon_t^s(z)$, is assumed serially uncorrelated, normally distributed with zero mean and variance σ_ϵ^2. This variance is assumed to be equal in all markets. For each z, $\alpha^s(z)$ is the short-run relative price elasticity of supply. Disparity in the supply elasticities of different goods follows from heterogeneous production functions. However, in the long run relative prices are assumed

fixed because of perfect substitutability on the supply side. The long run is measured here by one period, after which all suppliers can shift to other markets.

Looking one period ahead, all the markets offer the same mean price, but, as shown below, the corresponding variances differ according to the excess demand elasticities. Because agents are risk neutral, they are indifferent between the markets, and thus they choose a market for the next period randomly. There is an additional point related to the ex ante variability of the individual prices. Intuitively, one would expect a market with more price variance to be less desirable because local information would yield price level estimates of lower precision. However, as shown below, this turns out not to be the case.

On the demand side, the relative price elasticities are assumed constant across markets. The demand function also includes the term $M_t - EP_t$, which accounts for a real balance effect.

At the beginning of each period the stock of money in the economy is increased by transfers from the government to the public. This new money is assumed to be distributed equally across the markets. Within each market, however, the transfers are allocated randomly among a large number of agents. The rate of growth of the money stock, $m_t = M_t - M_{t-1}$, obeys

$$m_t = \sum_i \beta_i X_{it} + \tilde{m}_t \equiv g_t + \tilde{m}_t,$$

where the X_{it}'s are variables (past or current) that can be observed in all locations and the β's are known coefficients. The quantity $\tilde{m}_t$ is a random variable with zero mean and variance σ_m^2. Thus, g_t is the expectation about money growth formed from all the economy-wide shared information. It can be considered the prior expectation. The posterior is formed using the additional information conveyed by the local price. Thus, while g_t is the same everywhere, the posterior expectation Em_t is conditional on location-specific information as well and therefore varies across markets.

From equations (1) and (2) market clearing implies that

$$P_t(z) = \{1 - 1/[\alpha^s(z) + \alpha^d]\} EP_t + \{1/[\alpha^s(z) + \alpha^d]\} [M_t + \epsilon_t(z)]. \quad (3)$$

For each z, the sum $\alpha^s(z) + \alpha^d$ is the relative price elasticity of excess demand. Let $\lambda(z) \equiv 1/[\alpha^s(z) + \alpha^d]$. Each market has a constant $\lambda(z)$, but across markets $\lambda(z)$ is distributed according to a given density function with average value λ and "variance" σ_λ^2. Consistent with the assumption that agents possess accurate knowledge about the structure of the economy, this distribution is assumed to be known.

Following Lucas (1973) and Barro (1976), the solution for prices in terms of exogenous variables is obtained using the method of unde-

termined coefficients. Given the log-linearity of the model, the solution for the aggregate price level has the form

$$P_t = \Pi_1 M_{t-1} + \Pi_2 g_t + \Pi_3 \tilde{m}_t. \tag{4}$$

Namely, the aggregate price level will be related to the current money stock, which is divided into its different components. Lagged values, if added to (4), yield zero coefficients. Since M_{t-1} and g_t are fully perceived at time t, taking the expectation of both sides yields

$$EP_t = \Pi_1 M_{t-1} + \Pi_2 g_t + \Pi_3 E\tilde{m}_t. \tag{5}$$

The conditional expectation of m_t is now computed. Rewrite (3) as:

$$\delta_t(z) = g_t + \tilde{m}_t + \epsilon_t(z),$$

where

$$\delta_t(z) = [1/\lambda(z)]P_t(z) - [1/\lambda(z) - 1]EP_t - M_{t-1}.$$

The total disturbance affecting market z, $\delta(z)$, is partly nominal and partly real. Agents perceive $\delta(z)$ and form their expectations about its components. Given the stochastic specification of m_t and $\epsilon_t(z)$, the mean of the distribution of m_t conditional on $\delta(z)$ is

$$Em_t = g_t + \frac{\sigma_m^2}{\sigma_m^2 + \sigma_\epsilon^2}[\delta_t(z) - g_t],$$

or

$$Em_t = g_t + \frac{\sigma_m^2}{\sigma_m^2 + \sigma_\epsilon^2}[\tilde{m}_t + \epsilon_t(z)]. \tag{6}$$

Observe that $\lambda(z)$ does not appear in (6). Since agents located in z know this elasticity, they are able to isolate the composite disturbance independently of $\lambda(z)$. Thus, while the ex ante variance of prices depends on the particular elasticity (this follows from eq. [26] below), the precision obtainable from the local information is independent of $\lambda(z)$. Indeed $P_t(z)$ would convey less valuable information in higher price variance markets if the differential variability was due to a disparity in σ_ϵ^2. In this model, however, σ_ϵ^2 is the same across markets.

Substitute now (6) into (5) and the resulting expression for EP_t into (3) to obtain

$$EP_t = \Pi_1 M_{t-1} + \Pi_2 g_t + \Pi_3 \frac{\sigma_m^2}{\sigma_m^2 + \sigma_\epsilon^2}[\tilde{m}_t + \epsilon_t(z)], \tag{7}$$

$$\begin{aligned} P_t(z) = [1 - \lambda(z)]\left\{\Pi_1 M_{t-1} + \Pi_2 g_t + \Pi_3 \frac{\sigma_m^2}{\sigma_m^2 + \sigma_\epsilon^2}[\tilde{m}_t + \epsilon_t(z)]\right\} \\ + \lambda(z)[M_{t-1} + g_t + \tilde{m}_t + \epsilon_t(z)]. \end{aligned} \tag{8}$$

A new expression for the general price level can be computed from (8) by averaging with respect to the densities of $\lambda(z)$ and $\epsilon_t(z)$:

$$P_t = (1-\lambda)\left(\Pi_1 M_{t-1} + \Pi_2 g_t + \Pi_3 \frac{\sigma_m^2}{\sigma_m^2 + \sigma_\epsilon^2}\tilde{m}_t\right) + \lambda(M_{t-1} + g_t + \tilde{m}_t). \tag{9}$$

Since equation (9) is identical to (4), the solution for Π_1, Π_2, and Π_3 is obtained by equating the corresponding coefficients in the two equations:

$$\begin{aligned} \Pi_1 &= 1, \\ \Pi_2 &= 1, \\ \Pi_3 &= \frac{\sigma_m^2 + \sigma_\epsilon^2}{\sigma_m^2 + (1/\lambda)\sigma_\epsilon^2}. \end{aligned} \tag{10}$$

Substituting (10) into (8) and (9) and rearranging terms yields the solution for the individual commodity price and the average price level:

$$P_t(z) = M_{t-1} + g_t + \frac{\sigma_m^2 + \lambda(z)(1/\lambda)\sigma_\epsilon^2}{\sigma_m^2 + (1/\lambda)\sigma_\epsilon^2}[\tilde{m}_t + \epsilon_t(z)], \tag{11}$$

$$P_t = M_{t-1} + g_t + \frac{\sigma_m^2 + \sigma_\epsilon^2}{\sigma_m^2 + (1/\lambda)\sigma_\epsilon^2}\tilde{m}_t. \tag{12}$$

The resulting actual relative price is:

$$P_t(z) - P_t = (1-\theta)\tilde{\lambda}(z)\tilde{m}_t + [\theta + \lambda(z)(1-\theta)]\epsilon_t(z), \tag{13}$$

where $\tilde{\lambda}(z) \equiv \lambda(z) - \lambda$ and $\theta \equiv \sigma_m^2/[\sigma_m^2 + (1/\lambda)\sigma_\epsilon^2]$.

The hypothesis expressed by equation (13) is that only the unperceived part of money growth can affect price relationships. Note that the realized values of the unperceived money growth appear in the relative price expression. This follows from the confusion between $\tilde{m}_t$ and $\epsilon_t(z)$. Since in general $E\tilde{m}_t \neq \tilde{m}_t$, part of the money shocks is mistakenly perceived to be a shift in relative excess demand. The ensuing short-run supply reactions differ across markets according to $\alpha^s(z)$, thus causing dispersion among actual prices. On the other hand, g_t is correctly identified as an aggregate disturbance and therefore cannot be confused with a relative shift of excess demand. The neutrality of perceived money follows from $y_t^s(z)$ and $y_t^d(z)$ being functions of the relative price and from the one-to-one relationship between g_t and the expected price level (eqq. [7] and [10]). Given some value for g_t, the quantities along the supply-and-demand schedules are the same as before, for local nominal prices higher by an amount equal to

the adjustment in EP_t—which equals g_t. Therefore, the market clears at a $P_t(z)$ which is higher by the same degree in all markets.[2]

The variance of relative prices at time t, defined as $\tau_t^2 \equiv 1/N \sum_{z=1}^N [P_t(z) - P_t]^2$, where N is the "very large" total number of markets in the economy, can be computed from equation (13)[3]

$$\tau_t^2 = \{(1-\theta)^2\sigma_\lambda^2 + [\theta + \lambda(1-\theta)]^2\}\sigma_\epsilon^2 + (1-\theta)^2\sigma_\lambda^2\tilde{m}_t^2. \tag{14}$$

An empirical test of this equation requires a measure of dispersion among prices or price indexes of different commodities. Mills (1927, chap. 3) discusses problems that the interpretation of this dispersion measure presents. For example, long-run differential technological changes will cause prices to disperse over time. One would like to filter out such effects, because the focus here is on short-run distortions caused by incomplete current information. The problem is alleviated by using rates of price *change* rather than price levels. Different trends do not affect the variation of price change dispersion over time—although alterations in these trends will. Thus, some of the long-run relative prive movements effects can be filtered from the dispersion measure. What remains can be considered to be captured by the random term in the dispersion equation.

The variance of the rates of change in individual prices is calculated using equation (13) and the equivalent for $t-1$. This variance, defined as

$$\gamma_t^2 \equiv \frac{1}{N}\sum_{z=1}^{N} \{[P_t(z) - P_{t-1}(z)] - (P_t - P_{t-1})\}^2, \tag{15}$$

follows as

$$\gamma_t^2 = 2(1-\theta)^2\sigma_\lambda^2\sigma_\epsilon^2 + 2[\theta + \lambda(1-\theta)^2]\sigma_\epsilon^2 + (1-\theta)^2\sigma_\lambda^2(\tilde{m}_t - \tilde{m}_{t-1})^2.$$

Equation (15) is the final price dispersion equation that is generated

[2] In order to consider the effects of changes in the velocity of money circulation on price dispersion, I worked out a similar simple model in which there is some current public information about future money growth. This information may be conveyed by political or military events that are believed to have implications for the future state of government finances. The prediction of future monetary expansion, which generates inflationary expectations, affects the velocity of circulation in the current period. With respect to relative prices, if the knowledge about future money growth is shared economy-wide, they will be unaffected by the change in velocity. This neutral effect follows from the same mechanism determining the neutrality of perceived money. Since all agents share the same knowledge and are assumed to use it in the same model to predict its effects, they will equally adjust their EP_t according to the change in velocity taking place. A "one-time jump" in all prices therefore occurs, without affecting their dispersion.

[3] For this computation, since $\epsilon_t(z)$ and $[\epsilon_t(z)]^2$ are independent of $\lambda(z)$, $[\lambda(z)]^2$, and $\tilde{\lambda}(z)$, the following equalities are used: $(1/N)\Sigma\epsilon_t(z)\cdot\lambda(z) = 0\cdot\lambda = 0$, $(1/N)\Sigma[\epsilon_t(z)]^2[\tilde{\lambda}(z)]^2 = \sigma_\epsilon^2\sigma_\lambda^2$, and $(1/N)\Sigma[\epsilon_t(z)]^2[\lambda(z)]^2 = \sigma_\epsilon^2(\sigma_\lambda^2 + \lambda^2)$.

by the model. Because it deals with the dispersion of price changes, the appropriate monetary-shocks variable is the magnitude of changes in $\tilde{m}_t$.

Consider next the implied relationship between the variance of money shocks and the dispersion of relative prices. Barro's theoretical result was that σ_m^2 is positively correlated with relative price variability. However, the effect of σ_m^2 is ambiguous in this extended version of the model, since it has different and opposite effects on the three terms in the γ_t^2 expression. The second term on the right-hand side of (15) is the remainder of the expression when all markets are alike; that is, when, as in Barro's case, all have the same excess demand elasticity ($\sigma_\lambda^2 = 0$). This term corresponds to his relative price variance, which depends positively on σ_m^2 when $0 < \lambda < 1$. This condition is the counterpart to Barro's assumption that substitution effects dominate wealth effects.

The first term accounts for the positive interaction between the diversity in elasticities and the strength of the relative shifts. A term of this sort would be included also in the dispersion expression under full current information. In the present case of partial information, the fraction $(1 - \theta)$ appears here because agents typically underestimate the magnitude of the relative shifts, thus diminishing their effect on price dispersion. Because this underestimation increases with σ_m^2, the first term is negatively related to the money variance. The other, more interesting negative effect of σ_m^2 appears in the third term, namely, in the coefficient of $(\tilde{m}_t - \tilde{m}_{t-1})^2$. If σ_m^2 increases—or more precisely when the public perceives it doing so—money disturbances are less confused with real shifts, implying that a given shock induces smaller dispersion. This effect is a relative price equivalent of Lucas's hypothesis about the link between the variance of the nominal disturbances and the slope of the Phillips curve.

In the testing of equation (15), reported in Section IV, an attempt is made to capture the different effects of σ_m^2 and its net influence on price dispersion. However, the procedure adopted does not indicate that shifts in σ_m^2 have an important effect.

II. Construction of the Price Dispersion Series

This section reports the computation of a measure of price dispersion for the hyperinflation in Germany during the period January 1921–July 1923. The data set, consisting of 68 series of monthly averages of wholesale commodity prices, is obtained from the German statistical yearbook issues of 1921/22 and 1923 (see *Statistisches Reichsamt* 1921/22 and 1923). Other series from this source, some reported only until December 1921 (seven commodities) and others beginning only in

January 1922 (21 commodities), were deleted in order not to introduce a bias due to changes in the sample size and composition.

Prices are quoted from commodity exchanges of several German cities.[4] Each series, however, originates in a single location. The 68 commodities include 27 foodstuffs, 19 textiles and leathers, and 22 metals, oils, and coals. They are not finished goods but materials in a rather raw state. Because weights for the different commodities are unfortunately not available, unweighted rates of price change are used. Hopefully, the wide range of commodities in the sample approximates the general relative price instability during that period.

The individual price rates of change are computed as the first difference of the logarithms of the prices. Average values and variances are then calculated using

$$\Delta P_t = \frac{1}{N} \sum_i \Delta P_{it},$$

$$\gamma_t^2 = \frac{1}{N} \sum_i (\Delta P_{it})^2 - (\Delta P_t)^2,$$

where P_{it} is the price of commodity i and $\Delta P_{it} = \log P_{it} - \log P_{it-1}$. Table 1 contains the computed values of ΔP_t and γ_t^2. Due to missing observations, the actual number of commodities included in the calculations varies slightly from month to month. The third column in table 1 indicates the number of commodities for which both P_{it} and P_{it-1} are available.

III. Estimation of the Unperceived Part of Money Growth

Determination of the unperceived component of money growth during the hyperinflation requires a specification of the information set assumed to have been available to the public and the functional form for calculating the conditional expectation of money growth. Consider the expectation conditioned on economy-wide or "global" information g_t. This term was defined in Section I to be the prior expectation and is distinguished from the posterior expectation because it does not incorporate the additional information derived from local price observations.

This global information is assumed to consist of the current government spending in foreign exchange units, S_t, the current exchange rate, e_t, and 1-month lagged data on the money stock, price level, and all other macroeconomic variables. Not included is government reve-

[4] Given this source, these data do not present the problem of reported wholesale price data in the United States, discussed by Stigler and Kindahl (1970), that they do not always reflect discounts from list prices.

TABLE 1

MEAN AND VARIANCE OF WHOLESALE RATES OF PRICE CHANGES, GERMANY, FEBRUARY 1921–JULY 1923

Month	ΔP_t	γ_t^2	Number of Commodities
1921:			
February	−.09	.015	63
March	−.05	.011	66
April	−.02	.007	66
May	−.01	.026	66
June	.05	.032	67
July	.04	.034	66
August	.19	.081	66
September	.19	.033	68
October	.24	.034	68
November	.38	.043	68
December	−.03	.062	66
1922:			
January	.05	.018	66
February	.12	.013	68
March	.25	.011	65
April	.12	.018	65
May	.04	.013	66
June	.09	.007	66
July	.36	.017	67
August	.70	.097	68
September	.40	.071	66
October	.68	.064	66
November	.78	.038	66
December	.22	.052	66
1923:			
January	.70	.064	66
February	.69	.099	66
March	−.17	.041	65
April	.11	.018	65
May	.50	.039	66
June	.82	.048	62
July	1.25	.151	63

SOURCE.—Based on monthly average price data from *Statistisches Jahrbuch für das Deutsche Reich* (1921/22), pp. 282–83, and (1923), pp. 286–89.

nue from taxation and other sources because this variable depends on the current level of economic activity and is unlikely to be preannounced and to be widely known contemporaneously. It is natural to assume that the part of government expenditure consisting of the reparations to the Allied Powers was known in foreign exchange terms. With respect to the other expenditures, the implication is that nominal spending was observable and could be readily converted given the exchange rate.[5]

[5] A referee of this *Journal* suggested an alternative procedure for isolating unper-

The prior expectation of money growth is derived from the government monthly budget constraint, namely,

$$M_t^0 - M_{t-1}^0 = S_t^0 e_t^0 - \text{(other forms of nominal government revenue)}. \tag{16}$$

The superscript 0 indicates that the variables are not in logs but in their original form. Equation (16) indicates that creation of high-powered money equals the part of nominal expenditure that is not financed in some other way. The expression M_t^0 would correspond here to the end-of-month money stock. The other forms of government finance are taxes, net sale of bills, gold sold to the public, etc.

If this other revenue comprised an approximately fixed proportion of total expenditure over time, the budget equation above could be expressed as

$$M_t^0 - M_{t-1}^0 = kS_t^0 e_t^0 + \text{random term}, \tag{17}$$

where k $(0 < k < 1)$ is the average fraction of the expenditure financed by money issue.

The first attempt to generate a perceived money growth series was made using an equation of this type. Dividing (17) through by M_{t-1}^0, money growth appears linearly related to $S_t^0 e_t^0 / M_{t-1}^0$. The three variables in this ratio are assumed currently known, and, therefore, this specification is consistent with the notion that the conditional expectation can be formed using only currently observable variables.

However, a regression of this form,[6] including a constant, shows that k was probably not constant over the period. Specifically, the existence and pattern of residual serial correlation,[7] plus some additional considerations discussed below, suggest a nonlinear relationship between money issue and spending during that period.

Assuming, then, that the fraction k is not constant over time, the question is whether something can be said about its determinants. In

---

ceived money using information from the Central Bank's balance sheet. This information is available monthly in Statistisches Reichsamt (1925). When the Central Bank discounted a bill, the corresponding credit appeared in the current account of the state or the private borrower. Only when these funds were withdrawn did the notes in circulation increase. Using information about these credits, agents could forecast money growth in the near future.

[6] The estimated OLS equation is the following (see below for the inclusion of the lagged spending variable):

$$\frac{M_t^0 - M_{t-1}^0}{M_{t-1}^0} = \underset{(.019)}{-.049} + \underset{(.033)}{.317}\,\frac{S_t^0 e_t^0}{M_{t-1}^0} + \underset{(.089)}{.381}\,\frac{S_{t-1}^0 e_{t-1}^0}{M_{t-1}^0}$$

$$(R^2 = .95,\ \text{D-W} = 1.0,\ \sigma = .067).$$

[7] The residuals are generally negative at the beginning and the end of the period, approximately the low and high values of money growth, and generally positive for the rest of the sample.

order to suggest an answer to this question, rewrite equation (17) as

$$\frac{M_t^0 - M_{t-1}^0}{M_t^0} = k_t \frac{S_t^0}{M_t^0/e_t^0} + \text{random term}. \tag{18}$$

Equation (18) preserves the positive correlation between money growth and the ratio of real expenditure to real cash balances, but, unlike (17), the fraction k is now allowed to vary over time. It is now argued that k_t is itself correlated with M_t^0/e_t^0 and S_t^0.

To examine this correlation, assume first that S_t^0 is fixed at some value S^0. This level of real spending can be financed by different mixes of inflationary finance on the one hand, and taxation, debt issue, etc., on the other, where the amount to be collected by money issue is expressed as k_tS^0. In the usual diagram plotting the demand for real balances as a function of the inflation rate, k_tS^0 is measured in steady states by the area of the rectangle defined by μ—the inflation rate—and M_t^0/e_t^0.

Consider now an increment in μ. Real balances decline according to the money demand function, and the revenue from inflation, k_tS^0, increases as long as μ is below the rate that corresponds to a unitary demand elasticity for real balances. Because real spending is constant, k_t increases, and, hence, the fraction of S^0 financed by other means declines.

This shift from taxation to money issue can be viewed as the policy variable that brings about higher inflation rates. Classic works on the German hyperinflation, like those by Graham (1930) and Bresciani-Turroni (1937), describe an opposite direction of effect. Namely, the rate of depreciation of the currency had a negative effect on the real yield from taxation due to the interval of time existing between the occurrence of taxable transactions and the actual payment of the taxes.[8] The present discussion relies on the correlation between the fraction of expenditure financed by money issue and the inflation rate rather than on a specific mechanism relating these two variables. This positive correlation implies that k_t and $1/(M_t^0/e_t^0)$ move in the same direction. However, this coincidental movement does not hold for all μ. When μ reaches the rate that maximizes the revenue from inflation, k_t also reaches its highest level, and when it rises above that rate, k_t must decline. In other words, the correlation between k_t and $1/(M_t^0/e_t^0)$ turns negative in that range.

This decline in k_t implies that the revenue from other sources must go up. If tax collection and debt issue cannot be increased (e.g., due to the negative effect of inflation mentioned above), spending must be partially financed by extraordinary means, such as sales of gold from

[8] See, e.g., Bresciani-Turroni 1937, p. 66; and Graham 1930, p. 44.

the Central Bank's stock. In fact, the balance sheet of the German Central Bank shows that the stock of gold begins to decline significantly in April 1923, after being fairly stable since 1920.[9]

Given this behavior of k_t when S_t^0 is constant, equation (18) can be approximated by the semilogarithmic form:

$$\frac{M_t^0 - M_{t-1}^0}{M_t^0} = \text{constant} + b' \log\left(\frac{1}{M_t^0/e_t^0}\right) + u', \tag{19}$$

where b' is a positive coefficient, u' is a random term of zero mean, and the constant term is affected by the level of S^0. In this specification, the implicit fraction k_t increases along with $1/(M_t^0/e_t^0)$ at lower and middle ranges of this variable but eventually declines when real balances fall below a certain value.

Equation (19) acquires more empirical content if real spending is allowed to vary. During the period under study, S_t^0 has a declining trend. However, holding constant the real balances term, which has a strong correlation with time, movements in S_t^0 can be interpreted as temporary deviations from a "normal" trend. These fluctuations are assumed also to be correlated with the fraction k_t. The assumption here is that given relatively high costs associated with temporary shifts in tax collection and debt issue, transitory movements in spending would be financed primarily by adjustments in money issue. A positive correlation between S_t^0 and k_t would then result. However, a sufficiently high value of S_t^0 could be presumed to require extraordinary finance of the sort previously mentioned, so that k_t might eventually decline.

Incorporating an approximation of this effect into equation (19) results in the following generalized expression:

$$\frac{M_t^0 - M_{t-1}^0}{M_t^0} = a' + b' \log\left(\frac{1}{M_t^0/e_t^0}\right) + c' \log S_t^0 + u',$$

which can be rewritten as

$$\frac{M_t^0 - M_{t-1}^0}{M_t^0} = a' - b'[M_{t-1} + (M_t - M_{t-1}) - e_t] + c'S_t + u', \tag{20}$$

where variables without a superscript are again in logarithmic terms.

In order to proceed with the formulation of the prior expectation, it is convenient to replace the logarithmic growth rate $(M_t - M_{t-1})$ on the right-hand side by the growth rate measured by $(M_t^0 - M_{t-1}^0)/M_t^0$. While this rate is always lower than $M_t - M_{t-1}$, the gap widening the higher the growth rates, this effect can hopefully be captured approximately by the coefficients in the estimated equation. Then,

[9] See Statistisches Reichsamt 1925, p. 53.

equation (20) can be solved for $(M_t^0 - M_{t-1}^0)/M_t^0$ to yield

$$\frac{M_t^0 - M_{t-1}^0}{M_t^0} = \frac{a}{1+b} - \frac{b}{1+b}(M_{t-1} - e_t) + \frac{c}{1+b}S_t + \frac{1}{1+b}u_t. \quad (21)$$

The prior conditional expectation is defined accordingly as

$$g_t \equiv \left(\frac{\hat{a}}{1+b}\right) - \left(\frac{\hat{b}}{1+b}\right)(M_{t-1} - e_t) + \left(\frac{\hat{c}}{1+b}\right)S_t. \quad (22)$$

The unperceived part of money growth $\tilde{m}_t$[10] is then computed by the difference between actual growth and g_t, namely,

$$\tilde{m}_t = (M_t^0 - M_{t-1}^0)/M_t^0 - g_t.$$

The coefficients in equation (22) are those which result from regressing $(M_t^0 - M_{t-1}^0)/M_t^0$ on S_t and $(M_{t-1} - e_t)$. However, the exchange rate is not in general an exogenous variable in a money growth equation. A correlation between e_t and the error term u_t will exist via some unspecified condition for equilibrium in the foreign exchange market. Therefore, the coefficients in (22) do not correspond exactly to those in equation (21). This property is not a drawback. On the contrary, the bias in the estimated coefficients (relative to those in [21]) reflects the part of u_t that can be estimated from e_t. It therefore should be taken into account in calculating g_t.[11]

Turn now to the estimation of equation (21). There is a problem in matching the available data on money with those on prices for the German hyperinflation. Unlike the price series, which consist of monthly averages, the available data on the money stock until January 1923 are end-of-month figures.[12] From January 1923 onward, four quotations per month are available. Thus, a proxy is constructed for the monthly average money stock. For the period January–July 1923 it contains averages of the beginning-of-month, end-of-month, and the three intermediate quotations available. Until December 1922, the monthly averages are approximated by linear interpolation of the end-of-month figures. The constructed series are shown in table 2.

A further consideration arises. The estimation of equation (21) from monthly average data on the money stock, rather than end-of-

[10] The unperceived growth $\tilde{m}_t$ does not correspond exactly to the error term u_t. On this point, see below.

[11] E.g., in the general linear model $y = X\beta + u$, where the variables in X are correlated with u, the estimated vector of coefficients is $\hat{\beta} = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + u)$, or $\hat{\beta} = \beta + (X'X)^{-1}X'u$, where $(X'X)^{-1}X'u$ is the regression coefficient of u on X. The prediction of y_t given the values of the vector x_t is, accordingly, $y_t = x_t'[\beta + (X'X)^{-1}X'u]$; i.e., it is composed of the systematic part $x_t'\beta$, plus the conditional expectation of u_t given x_t.

[12] Another problem stems from the form of the prior expectation, which, as it stands, requires the use of end-of-month money stocks. See the discussion below.

TABLE 2

VALUES (in Logarithms) OF REAL EXPENDITURE, EXCHANGE RATE, AND MONEY STOCK

	S_t	e_t	M_t
Oct. 1920	5.85	2.79	11.24
Nov.	6.44	2.91	11.25
Dec.	6.58	2.86	11.28
Jan. 1921	6.42	2.74	11.29
Feb.	6.80	2.68	11.28
Mar.	6.66	2.70	11.29
Apr.	6.65	2.71	11.30
May	6.53	2.69	11.30
June	6.76	2.80	11.33
July	6.43	2.91	11.36
Aug.	6.76	3.00	11.38
Sept.	6.23	3.22	11.42
Oct.	5.95	3.58	11.48
Nov.	5.54	4.14	11.55
Dec.	6.44	3.82	11.66
Jan. 1922	5.96	3.82	11.72
Feb.	5.84	3.90	11.75
Mar.	5.85	4.22	11.81
Apr.	5.77	4.24	11.88
May	5.93	4.24	11.96
June	5.75	4.33	12.05
July	5.34	4.77	12.16
Aug.	5.32	5.60	12.33
Sept.	6.08	5.86	12.58
Oct.	5.60	6.63	12.92
Nov.	5.30	7.44	13.35
Dec.	6.08	7.50	13.85
Jan. 1923	5.32	8.36	14.26
Feb.	5.63	8.80	14.82
Mar.	6.55	8.53	15.31
Apr.	6.15	8.67	15.60
May	5.65	9.34	15.83
June	6.21	10.17	16.29
July	6.16	11.34	17.14

NOTE.—S_t = log of the monthly government expenditure in millions of gold marks. Source: Bresciani-Turroni (1937), pp. 436–37. e_t = log of the monthly average exchange rate of the gold marks in paper marks. Source: Bresciani-Turroni (1937), p. 441. M_t = log of the monthly average money stock in millions of paper marks. Source: based on fixed days quotations from *Sonderhefte zur Wirtschaft und Statistic* (1925), pp. 45–47.

month figures, means that both the current month's spending and that of the previous month should be considered. Spending financed by money issue during the previous month increases one to one the current monthly average stock but has, in general, a weaker effect on the prior month's average stock.[13] To account for this effect, lagged

[13] This effect can be seen by considering first a case where each month's spending is spread evenly over the month, and, say, it is financed only by the issue of new money. In this case the money stock grows linearly at, in general, different rates within each monthly period. Then, if $M^o_{t-2,\mathrm{end}}$ denotes the money stock at the end of month $t-2$,

spending is incorporated into the framework of the semilogarithmic function in equation (21). First, define the variables S'_t and e'_t by $S'_t \equiv \log [\xi S^0_t + (1 - \xi)S^0_{t-1}]$ and $e'_t \equiv \log [\xi S^0_t + (1 - \xi)S^0_{t-1}]$. After substituting S'_t and e'_t for S_t and e_t in (21), ξ is estimated, simultaneously with the other coefficients in the equation, using a nonlinear maximum likelihood procedure under normally distributed errors.[14]

The estimated nonlinear equation is

$$\frac{M^0_t - M^0_{t-1}}{M^0_t} = \underset{(.071)}{.660} - \underset{(.007)}{.166}\,(M_{t-1} - e'_t) + \underset{(.014)}{.115}S'_t \tag{23}$$

$$(\underset{(.11)}{\hat{\xi} = .75}, R^2 = .98, \text{D-W} = 1.63, \sigma = .026, 33 \text{ observations}),$$

where the numbers in parentheses are the standard errors of the coefficients. With respect to the number of observations, the starting month was taken as November 1920 in order to test lagged effects of monetary shocks on price dispersions. According to the argument of note 13 above, the value .75 for ξ suggests a pattern of spending that is biased toward the beginning of the month. In order to proceed with the analysis on the more familiar ground of linear equations estimation, ξ is assumed henceforth to equal .75. Given this value, the standard errors of the other coefficients, linearly estimated, are not materially different from those obtained above. In order of their appearance in (23), they are .068, .005, and .014. The standard error of the regression is now 0.0256. The other regression statistics and the coefficients remain obviously the same.

The pattern of the residuals from equation (23), which are reported in table 4, suggests that their variance increased during the sample period as inflation progressed. A method similar to that proposed by Glejser (1969) is adopted to correct for the apparent heteroscedasticity by assuming a specific model for the variance of the error term. In this procedure the variance is postulated to be determined by a set of variables $\{z_i\}$ in the linear form:

$$\sigma^2_{mt} = \sum_i \omega_i z_{it}. \tag{24}$$

---

the monthly average for $t - 1$ equals $M_{t-2,\text{end}} + \frac{1}{2} S^0_{t-1}e^0_{t-1}$, and that corresponding to month t equals $M^0_{t-2,\text{end}} + S^0_{t-1}, e^0_{t-1} + \frac{1}{2} S^0_t e^0_t$. Thus, the increase in the monthly average from $t - 1$ to t equals $\frac{1}{2} S^0_{t-1}e^0_{t-1} + \frac{1}{2} S^0_t e^0_t$. Namely, spending evenly spread over each month would imply equal weights for current and lagged spending in the money growth equation. Alternatively, if spending is concentrated at the beginning of the month, the relative weight of lagged expenditure would be lower. At the extreme, e.g., if all spending is made only on the first day of each month, both $t - 1$ and t monthly averages increase equally by the amount of the $t - 1$ expenditure. In this case lagged spending does not belong in the money growth equation.

[14] This procedure is from the TSP Regression Package.

If the values of the true money shocks $\tilde{m}_t^*$ were available, one could use them as follows. Since the expectation of $\tilde{m}_t^{*2}$ is σ_{mt}^2, it follows that

$$\tilde{m}_t^{*2} = \sigma_{mt}^2 + v_t,$$

where v_t is of zero mean. Combining the last two equations yields

$$\tilde{m}_t^{*2} = \sum_i \omega_i z_{it} + v_t. \tag{25}$$

Estimates of the coefficients in equation (24) could be obtained by regressing $\tilde{m}_t^{*2}$ on the z_i variables. The heteroscedasticity problem is also present here, but it will be ignored in what follows.

The values of $\tilde{m}_t^*$ are, however, unknown; only the estimated residuals $\tilde{m}_t$ are available from the OLS money growth equation. Since $\tilde{m}_t^2$ converges to $\tilde{m}_t^{*2}$ asymptotically, the variance estimated using $\tilde{m}_t^2$ values obtained from small samples will be biased. This bias is neglected, as it is probably of relatively small importance.

In order to proceed with the implementation of this procedure the set of the z_i variables must be specified. The presumption is that the same variables used to explain the growth rates are also correlated with the variances. Thus, S_t', e_t', and M_{t-1} are candidates. The lagged squared residual $\tilde{m}_{t-1}^2$ is also included as an explanatory variable. It presumably captures the effect of serially correlated omitted variables and perhaps a direct correlation between σ_{mt-1}^2 and the current variance. The estimated variance equation is

$$\begin{aligned}\tilde{m}_t^2 = -\underset{(.003)}{.007} - \underset{(.00047)}{.00038}S_t' - \underset{(.00031)}{.00026}e_t' + \underset{(.00047)}{.00090}M_{t-1} - \underset{(.130)}{.281}\tilde{m}_{t-1}^2\end{aligned} \tag{26}$$

$$(R^2 = .57, \text{D-W} = 2.5, \sigma = .001, 33 \text{ observations}).^{15}$$

Nothing in this procedure for estimating the series of money variances guarantees that all the fitted values from equation (26) would be positive. Indeed, two of the fitted values have a negative sign. In order to use the estimated series as a measure of variances, these two negative values are replaced with the smallest positive value in the series. The series of the square roots of these estimates are reported in column 4 of table 4.

Using this series as weights for the corresponding observations, equation (23) is reestimated with the following results:

$$(M_t^0 - M_{t-1}^0)/M_t^0 = \underset{(.047)}{.700} - \underset{(.006)}{.154}(M_{t-1} - e_t') + \underset{(.012)}{.094}S_t' \tag{27}$$

$$(R^2 = .95, \text{D-W} = 1.8, \sigma = 1.01, 33 \text{ observations}).$$

[15] In order to estimate the 33 variances needed to reestimate eq. (34), the $\tilde{m}_t$ series were obtained from running the money growth equation after adding the additional observation of October 1920.

Observe that the coefficient of S_t' here is somewhat lower than in the OLS equation and that the coefficient of $(M_{t-1} - e_t')$ is somewhat higher. The general form of the equation is, however, robust to this transformation of the data.

The next step is to test the stability of the coefficients in equation (27) across two subperiods. Stability of the coefficients has particular relevance here. If the equation is approximately stable, it seems easier to assume that it was known from the beginning of the period and that perceptions about money growth were formed using the same equation during the entire sample.[16] The period is divided into an approximately nonaccelerating money supply period until May 1922 and an accelerating phase beginning in June 1922. An F-test applied to these two subperiods yields the statistic $F^3_{27} = 1.7$, with a corresponding 5 percent critical value of 3.0. Therefore, the hypothesis of stable coefficients across these two subperiods cannot be rejected at the 5 percent significance level.

A regression in which the coefficients of M_{t-1} and e_t' are unconstrained produces coefficients of similar magnitude for the two variables. The F-test for the linear constraint of equal coefficients yielded the statistic $F^1_{29} = 2.3$, where the 5 percent critical value is 4.2.

IV. Empirical Test of the Dispersion Equation

The tests of the price dispersion model in equation (15) are performed using the dispersion series computed in Section II and the unperceived monetary shocks as measured by the residuals in equation (27).

For convenience equation (15) is rewritten here

$$\gamma_t^2 = \left\{2(1-\theta)^2\sigma_\lambda^2 + 2[\theta + \lambda(1-\theta)]^2\right\}\sigma_\epsilon^2 + (1-\theta)^2\sigma_\lambda^2(\tilde{m}_t - \tilde{m}_{t-1})^2. \tag{15}$$

Significantly, this equation has a simple linear form—and can therefore be tested by an ordinary least-squares procedure—only under constant money and relative shocks variances. However, the analysis of money growth in the previous section suggested that σ_m^2 increased during the hyperinflation. If this is indeed the case, it would not be appropriate to test the model with a specification that relates price dispersion to money shocks with a constant coefficient.

Two different procedures are adopted here to deal with the possibility of a changing money variance. The first uses a linear approximation in which the variance of money—as measured by the $\hat{\sigma}_m^2$ series

[16] Estimation of a money growth equation in a similar context, using the entire sample (for U.S., 1941–73), was discussed and performed in Barro (1977).

from Section III—is kept constant by including it additively in the equation. As discussed in Section I, σ_m^2 has different and opposite effects on γ_t^2, and, therefore, on an a priori basis the coefficient of $\hat{\sigma}_m^2$ could take either sign. The other attempt is to estimate (15) as a nonlinear equation. The results of this procedure, reported later in the section, are quite poor.

The estimated equation in which $\hat{\sigma}_m^2$ is added linearly is:

$$\gamma_t^2 = \underset{(.010)}{.033} + \underset{(3.2)}{17.4}(\tilde{m}_t - \tilde{m}_{t-1})^2 - \underset{(8.6)}{15.8}\hat{\sigma}_{mt}^2 \tag{28}$$

$(R^2 = .59$, D-W $= 1.2$, $\sigma = .022$, 30 observations).

The monetary shocks appear to have considerable explanatory power for price dispersion. The coefficient of $\hat{\sigma}_{mt}^2$ is negative and, therefore, suggests a dominant Lucas-type effect of the money variance on price dispersion. That is, the degree of dispersion associated with given shocks diminishes the higher their variance. The explanatory power of $\hat{\sigma}_{mt}^2$, however, is fairly low; its coefficient is significantly different from zero at the 5 percent level but fails to be so at the 2.5 percent level.

Theoretically, the 1-month lagged money variance $\hat{\sigma}_{mt-1}^2$ belongs also in the equation. However, when included, its coefficient is insignificant with a t-statistic of .8.

The Durbin-Watson statistic indicates autocorrelated residuals, which may be caused by omitted real variables (like changes in the pattern of government spending,[17] in income distribution, etc.) that are serially correlated or by the fact that the $\tilde{m}_t$ variable includes an estimation error. In order to check whether the degree of significance of the estimated coefficients in (28) is affected by this autocorrelation, the equation is reestimated using the Cochrane-Orcutt technique. The results are quite similar to those obtained before:

$$\gamma_t^2 = \underset{(.008)}{.038} + \underset{(2.9)}{16.8}(\tilde{m}_t - \tilde{m}_{t-1})^2 - \underset{(9.8)}{18.5}\hat{\sigma}_{mt}^2 \tag{29}$$

$(R^2 = .67$, D-W $= 2.2$, $\sigma = .020$, 29 observations, $\hat{\rho} = \underset{(.17)}{.36})$.

Including $\hat{\sigma}_{mt-1}^2$ in this regression yielded a t-statistic of only 1.0 for its coefficient. The possibility of lagged effects of monetary shocks on price dispersion was explored by including the variable $(\tilde{m}_{t-1} - \tilde{m}_{t-2})^2$ in the equation, but the estimated coefficient was found statistically insignificant. The t-ratio was 0.3 in the OLS regression and 1.4 using the Cochrane-Orcutt technique.

[17] The magnitude of changes in the amount of real government spending, however, does not have any significant explanatory power.

The next step is to see whether the dispersion equation is stable over the entire period. In fact, the inclusion of the $\hat{\sigma}^2_{mt}$ variable is an attempt to control one source of instability in the coefficients. In order to carry out this test the sample is divided first after May 1922, estimating the equation separately for the two subperiods. This partition of the sample is the same one adopted previously to test the stability of the money growth equation. Then, the exercise is repeated, dividing the sample at the end of 1922. The aim of this partition is to see how the model performs after removing from the sample the 7 months of 1923, which had a much more unstable monetary growth.

The results of these regressions, reported in table 3, can be summarized as follows. When the sample is divided in May 1922 (first column), the results for the first subperiod are fairly weak. Both coefficients are insignificant in the OLS equation. Using the Cochrane-Orcutt technique the coefficient of $(\tilde{m}_t - \tilde{m}_{t-1})^2$ turns out significant at the 5 percent level, although that corresponding to $\hat{\sigma}^2_{mt}$ is still insignificant. During the second subperiod—from June 1922 to July 1923—the statistical performance of the equation is much stronger.

The second column reports the equations estimated for the periods through and after December 1922. Observe that the removal of the 1923 portion of the sample worsens the performance of the model as judged by the size of the t-ratios. However, the coefficient of the monetary shocks still remains quite significant.

Formal F-tests fail to reject the hypothesis of stable coefficients across the mentioned subperiods. When the sample is divided in May 1922, the resulting statistic is $F^3_{24} = 1.8$, while the 5 percent critical value is 3.0. Partitioning the sample at the end of 1922 yields the statistic of only 0.8.

The other approach adopted to test the dispersion equation was to treat it as a nonlinear relationship. When σ^2_m is changing over time, equation (15) generalizes to

$$\begin{aligned}\gamma_t^2 = {} & \sigma_\epsilon^2\sigma_\lambda^2\left[\left(\frac{\sigma_\epsilon^2}{\lambda\sigma_{mt}^2+\sigma_\epsilon^2}\right)^2+\left(\frac{\sigma_\epsilon^2}{\lambda\sigma_{mt-1}^2+\sigma_\epsilon^2}\right)^2\right] \\ & + \sigma_\epsilon^2\left[\left(\frac{\lambda\sigma_{mt}^2+\lambda\sigma_\epsilon^2}{\lambda\sigma_{mt}^2+\sigma_\epsilon^2}\right)+\left(\frac{\lambda\sigma_{mt-1}^2+\lambda\sigma_\epsilon^2}{\lambda\sigma_{mt-1}^2+\sigma_\epsilon^2}\right)\right] \\ & + \sigma_\lambda^2\left(\frac{\sigma_\epsilon^2}{\lambda\sigma_{mt}^2+\sigma_\epsilon^2}\,\tilde{m}_t-\frac{\sigma_\epsilon^2}{\lambda\sigma_{mt-1}^2+\sigma_\epsilon^2}\,\tilde{m}_{t-1}\right)^2.\end{aligned} \tag{30}$$

As discussed in Section I, the third term in this equation reflects the negative effect of σ^2_m on the impact of monetary shocks. The second term corresponds to Barro's (1976) relative price variance expression

TABLE 3

ESTIMATED PRICE CHANGE DISPERSION EQUATIONS

	Feb. 1921–May 1922	Feb. 1921–Dec. 1922	Feb. 1921–July 1923
OLS	$.030 + 18.3(\tilde{m}_t - \tilde{m}_{t-1})^2 - 28.2\hat{\sigma}^2_{mt}$ (.012) (15.8) (32.2) ($R^2 = .22$, D-W = 1.2, $\sigma = .019$)	$.024 + 34.4(\tilde{m}_t - \tilde{m}_{t-1})^2 - 7.3\hat{\sigma}^2_{mt}$ (.009) (12.1) (19.0) ($R^2 = .29$, D-W = 1.1, $\sigma = .022$)	$.033 + 17.4(\tilde{m}_t - \tilde{m}_{t-1})^2 - 15.8\hat{\sigma}^2_{mt}$ (.010) (3.2) (8.6) ($R^2 = .59$, D-W = 1.2, $\sigma = .022$)
Cochrane-Orcutt	$.035 + 27.7(\tilde{m}_t - \tilde{m}_{t-1})^2 - 43.1\hat{\sigma}^2_{mt}$ (.012) (13.5) (32.0) ($R^2 = .47$, D-W = 2.5, $\sigma = .016$)	$.041 + 30.7(\tilde{m}_t - \tilde{m}_{t-1})^2 - 35.1\hat{\sigma}^2_{mt}$ (.012) (9.0) (21.2) ($R^2 = .53$, D-W = 2.2, $\sigma = .018$)	$.038 + 16.8(\tilde{m}_t - \tilde{m}_{t-1})^2 - 18.5\hat{\sigma}^2_{mt}$ (.008) (2.9) (9.8) ($R^2 = .67$, D-W = 2.2, $\sigma = .020$)
	June 1922–July 1923	**Jan. 1923–July 1923**	
OLS	$.054 + 17.5(\tilde{m}_t - \tilde{m}_{t-1})^2 - 28.1\hat{\sigma}^2_{mt}$ (.012) (3.6) (11.6) ($R^2 = .70$, D-W = 1.6, $\sigma = .022$)	$.036 + 16.6(\tilde{m}_t - \tilde{m}_{t-1})^2 - 16.4\hat{\sigma}^2_{mt}$ (.026) (3.2) (15.0) ($R^2 = .87$ D-W = 3.1, $\sigma = .020$)	

that is affected positively by σ_m^2 when $0 < \lambda < 1$. Recall that λ is defined as the average of $1/[\alpha^s(z) + \alpha^d]$ across markets, and $\alpha^s(z) + \alpha^d$ is the excess demand elasticity of commodity z. Using again the $\hat{\sigma}_m^2$ series, the parameters of equation (30)—λ, σ_ϵ^2, and σ_λ^2—are estimated using a nonlinear least-squares procedure with fairly weak results. The additional structure given to the equation seems to be rejected by the data, as judged by a higher sum of squared errors than in the linear equation.

Observe that a value of zero for λ reduces the equation to the linear form of equation (15), in which the money variance is constant. The estimated value for this parameter was 0.056 with a standard error of 0.065.[18] The interpretation of this result is not that the average $1/[\alpha^s(z) + \alpha^d]$ is likely to be close to zero. Instead, it suggests that the detailed specification of equation (30) is too stringent. For example, if the variance of the relative excess demand shifts, σ_ϵ^2 also changes over time; this is more of a problem in this approach, since σ_ϵ^2 itself is being estimated as a constant.

V. Actual Money Growth and Inflation in the Dispersion Equation

In this section additional variables that were mentioned in the literature as being related to price dispersion are tried in the equation. There is no rigorous theoretical justification for their inclusion. Thus, only loose verbal explanations are given. Also, the variables related to inflation are clearly not exogenous, and, therefore, any observed correlation cannot imply causality.

Actual money growth and price dispersion.—A variable that can be considered exogenous and, if one assumption of Section I is violated, in principle also relevant for price dispersion is the actual money growth. Changes in the money stock can affect the dispersion of prices (even when perceived) if the new money is spread unevenly across the economy, thereby affecting relative demand in different sectors. This type of effect was discussed by Cairnes (1873) with respect to gold discoveries. In the framework of the present model this type of effect could be represented by changes in the relative excess demand variance σ_ϵ^2. Since here the focus is on price change dispersion, the corresponding variable in this context is the change in the growth rate or the degree of acceleration/deceleration in the money stock.

In order to test this sort of effect, and also to see whether ($\bar{m}_t$ −

[18] $\hat{\sigma}_\lambda$ was 17.7 with a standard error of 8.4 and $\hat{\sigma}_\epsilon^2$ was 0.0008, with a standard error of 0.0004.

$\tilde{m}_{t-1})^2$ is only a proxy for changes in actual money growth, price dispersion was regressed on $[(M_t^0 - M_{t-1}^0)/M_t^0 - (M_{t-1}^0 - M_{t-2}^0)/M_{t-1}^0]^2$ with the following results:

$$\underset{}{\gamma_t^2} = \underset{(.006)}{.033} + \underset{(.6)}{2.1}\left(\frac{M_t^0 - M_{t-1}^0}{M_t^0} - \frac{M_{t-1}^0 - M_{t-2}^0}{M_{t-1}^0}\right)^2$$

$$(R^2 = .33,\ \text{D-W} = 1.5,\ \sigma = .028,\ 30\ \text{observations}).$$

The monetary acceleration/deceleration variable has a statistically significant correlation with price dispersion. Remarkably, however, its explanatory power vanishes when $(\tilde{m}_t - \tilde{m}_{t-1})^2$ and $\hat{\sigma}_{mt}^2$ are also included in the regression. The equation including the three variables is

$$\gamma_t^2 = .034 + \underset{(4.0)}{16.6}(\tilde{m}_t - \tilde{m}_{t-1})^2 - \underset{(9.0)}{16.5}\hat{\sigma}_{mt}^2 +$$

$$\underset{(.77)}{.25}\left(\frac{M_t^0 - M_{t-1}^0}{M_t^0} - \frac{M_{t-1}^0 - M_{t-2}^0}{M_{t-1}^0}\right)$$

$$(R^2 = .59,\ \text{D-W} = 1.3,\ \sigma = .022,\ 30\ \text{observations}).$$

This result denies the existence of any effect of relative demands following the introduction of new money during this period. It supports the hypothesis that money affects relative prices only if it is currently unperceived.

Inflation and price dispersion.—In his analysis of price behavior during the hyperinflation, Graham describes a positive correlation between the acceleration of price level and the dispersion of prices. Mills's findings in his study of U.S. prices suggest that dispersion is positively correlated with both acceleration and deceleration of the price level.

If the acceleration/deceleration in the price level is related to an unperceived monetary expansion or contraction, the theory tested here predicts that the correlation mentioned above should be captured by a variable measuring unperceived money growth. To test whether this is the case here, the variable $(\mu_t - \mu_{t-1})^2$ is also included in the equation—where μ_t is the inflation rate from month $t - 1$ to month t computed from the wholesale price index (see table 4). The results suggest that there is a separate correlation between $(\mu_t - \mu_{t-1})^2$ and price dispersion. The equation estimated by OLS is

$$\gamma_t^2 = \underset{(.005)}{.032} + \underset{(3.0)}{18.6}(\tilde{m}_t - \tilde{m}_{t-1})^2 - \underset{(8.7)}{24.2}\hat{\sigma}_{mt}^2 + \underset{(.031)}{.074}(\mu_t - \mu_{t-1})^2$$

$$(R^2 = .67,\ \text{D-W} = .9,\ \sigma = .020,\ 30\ \text{observations}).$$

Given the low D-W statistic, the equation was reestimated by

TABLE 4

VALUES OF MONEY GROWTH AND INFLATION

	$(M_t^0 - M_{t-1}^0)/M_t^0$ (1)	$\widehat{(M_t^0 - M_{t-1}^0)/M_t^0}$ (2)	Col. 1 − Col. 2 (3)	$\hat{\sigma}_m$ (4)	g_t (5)	$\tilde{m}_t$ (6)	μ_t (7)
Nov. 1920	.010	.003	.008	.008*	.006	.004	.029
Dec.	.026	.025	.001	.011	.024	.002	−.047
Jan. 1921	.008	−.006	.014	.014	−.003	.011	.001
Feb.	−.011	.010	−.021	.010	.008	−.019	−.045
Mar.	.010	.009	.001	.008*	.008	.002	−.028
Apr.	.007	.005	.002	.012	.004	.003	−.009
May	.008	−.009	.017	.014	−.007	.015	−.014
June	.022	.019	.003	.009	.017	.005	.043
July	.029	.012	.017	.014	.012	.016	.044
Aug.	.022	.042	−.020	.009	.038	−.016	.294
Sept.	.045	.036	.009	.008	.036	.009	.075
Oct.	.057	.041	.016	.017	.046	.011	.174
Nov.	.069	.076	−.007	.018	.083	−.014	.328
Dec.	.100	.117	−.017	.013	.113	−.013	.021

Jan. 1922	.061	.065	−.003	.017	.067	−.006	.050
Feb.	.024	.037	−.013	.023	.044	−.021	.113
Mar.	.060	.074	−.014	.020	.079	−.019	.281
Apr.	.075	.072	.003	.021	.078	−.003	.158
May	.070	.071	−.001	.024	.076	−.006	.015
June	.088	.059	.028	.026	.066	.022	.085
July	.106	.066	.040	.025	.077	.029	.358
Aug.	.158	.161	−.002	.020	.166	−.008	.646
Sept.	.221	.260	−.038	.023	.251	−.029	.402
Oct.	.285	.310	−.025	.016	.300	−.015	.679
Nov.	.349	.347	.002	.027	.339	.010	.712
Dec.	.393	.372	.021	.032	.355	.038	.245
Jan. 1923	.339	.367	−.027	.036	.355	−.016	.636
Feb.	.431	.380	.050	.038	.367	.063	.696
Mar.	.385	.364	.021	.034	.342	.043	−.133
Apr.	.252	.275	−.024	.047	.261	−.009	.064
May	.205	.268	−.064	.050	.260	−.056	.450
June	.371	.399	−.028	.037	.378	−.007	.864
July	.572	.518	.054	.049	.488	.084	1.350

NOTE.—$\widehat{(M_t^o - M_{t-1}^o)/M_t^o}$ is the estimated value of money growth from eq. (23). The data used in the estimation are shown in table 2. $\hat{\sigma}_m$ = square root of the estimated value from eq. (26). The entries with an asterisk are those with a negative fitted value that were replaced by the smallest positive value in the series. g_t = estimated value from the weighted least-squares regression (eq. [27]), where the series $\hat{\sigma}_m$ are used as weights. $\tilde{m}_t \equiv (M_t^o - M_{t-1}^o)/M_t^o - g_t$. μ_t = first difference of the logs of the Wholesale Price Index. Data obtained from *Sonderhefte zur Wirtschaft und Statistic* (1925).

Cochrane-Orcutt:

$$\gamma_t^2 = \underset{(.008)}{.033} + \underset{(2.6)}{19.5}(\tilde{m}_t - \tilde{m}_{t-1})^2 - \underset{(9.0)}{25.3}\hat{\sigma}^2_{mt} + \underset{(.022)}{.077}(\mu_t - \mu_{t-1})^2$$

$$(R^2 = .77, \text{D-W} = 2.00, \sigma = .017, 29 \text{ observations}, \hat{\rho} = \underset{(.16)}{.49}).$$

Not only does $(\mu_t - \mu_{t-1})^2$ have a statistically significant correlation with price dispersion, but its inclusion in the equation also sharpens the performance of the original two variables. Possible explanations of this correlation could be related, for example, to income redistribution following from unanticipated inflation or substitution between money and certain commodities as stores of value, when their relative costs change.

Sheshinski and Weiss (1977) consider a model of a monopolistic firm in which costs involved in changing the price of the commodity produced generate discrete periodic price adjustments whose magnitude increases with the inflation rate. They suggest that, if the timing of these adjustments is independent across firms, higher inflation rates imply larger dispersion of individual price changes. Hence, this type of argument seems to imply a relationship between price change dispersion and μ_t^2.[19]

A role for μ_t^2 may be rationalized also on different grounds. Graham mentions the effect of the interaction between government regulations and inflation on the dispersion of prices. He refers to price controls, such as rent restriction legislation, which generate dispersion among the differently affected prices as the general price level increases. According to this argument, the dispersion of prices will be positively correlated with the price level. Hence, the dispersion of price changes will increase with the magnitude of changes in the price level—namely with μ_t^2.

Adding μ_t^2 to the equation yields the following results:

[19] However, the empirical implications of Sheshinski and Weiss's analysis for price dispersion do not seem clear to me. An ambiguity arises because of the probably positive effect of inflation on the frequency of price changes that they derived. If the length of the observation period is kept constant, a higher frequency of price change may diminish the measured dispersion of price changes. The possibility of a negative effect of inflation on price dispersion in this framework can be seen in the following example. Assume that the optimal frequency of price adjustments for all firms is 2 months and that part of the firms adjust their prices during odd-numbered months and the rest during even-numbered months. Using monthly data, dispersion of price changes will depend on the magnitude of price changes corresponding to the group of firms currently adjusting prices. Now assume that inflation increases; as a consequence the magnitude of price adjustments goes up, and also the optimal frequency is increased—say to one per month. Since now all the firms adjust prices during the same month, the dispersion of price changes collapses to zero, in spite of the larger individual price changes.

$$\underset{(.005)}{\gamma_t^2 = .031} + \underset{(4.1)}{12.2}(\tilde{m}_t - \tilde{m}_{t-1})^2 - \underset{(8.2)}{21.3}\hat{\sigma}_m^2 + \underset{(.030)}{.062}(\mu_t - \mu_{t-1})^2 + \underset{(.017)}{.036}\mu_t^2$$

$$(R^2 = .67, \text{D-W} = 1.3, \sigma = .019, 30 \text{ observations}).$$

The coefficients of μ_t^2 and $(\mu_t - \mu_{t-1})^2$ are both significant at the 2.5 percent level. However, when the equation is reestimated by Cochrane-Orcutt, the effect of μ_t^2 becomes insignificant (t-ratio of 0.8), while the performance of the rest of the variables improves. If the monetary acceleration/deceleration variable is also included, its coefficient turns out negative but insignificant at the 5 percent level (t-ratio of 1.62 with OLS and 1.55 with Cochrane-Orcutt), without materially affecting the other coefficients and standard errors.

VI. Summary and Conclusions

This paper tested a model of price dispersion using data on the German hyperinflation after World War I. The theoretical framework, outlined in Section I, predicted that money growth causes price dispersion only if it is currently unperceived.

In order to perform this test, the first task was to construct a measure of price dispersion based on a broad enough range of commodities to approximate the varying degree of relative price variability during that period. This calculation was accomplished using wholesale price data on a sample of over 60 commodities, including foodstuffs, textiles, metals, and fuels.

The next step was the delicate one of measuring the perceived part of money growth. This estimation implied the postulation of an information set available at each point in time and also a function relating the variables in this set to the creation of money. The function used to isolate the perceived part of money growth was specified based on considerations related to the government demand for revenue to finance expenditure.

A measure of the variance of the unperceived money growth was also estimated. This series was used for dealing with the heteroscedasticity problem in the money growth equation and also as a theoretically relevant factor affecting price dispersion.

The testing of the dispersion equation showed a statistically significant correlation between unperceived money—as measured here—and price dispersion. Thus, the model is supported to some extent by the empirical results. The variance of the money shocks turned out to have a negative coefficient which was only marginally significant. Although the money variance has different and opposite theoretical effects on price dispersion, this negative correlation is consistent with a dominant effect of diminishing the misperceptions of

monetary disturbances as relative shifts—and thus reducing the dispersion associated with those disturbances.

In Section V actual money growth was also considered as a price dispersion factor, justified as affecting relative demands. However, the results indicated no additional effect for the rate of money creation. The magnitude of changes in the inflation rate does appear to have a separate correlation with price change dispersion. The present model, however, does not provide an explanation for this association.

References

Barro, Robert J. "Rational Expectations and the Role of Monetary Policy." *J. Monetary Econ.* 2 (January 1976): 1–32.

———. "Unanticipated Money Growth and Unemployment in the United States." *A.E.R.* 67 (March 1977): 101–15.

Bresciani-Turroni, Costantino. *The Economics of Inflation.* London: Allen & Unwin, 1937.

Cairnes, John E. "The Course of Depreciation." In *Essays in Political Economy: Theoretical and Applied.* London: Macmillan, 1873.

Glejser, H. "A New Test for Heteroskedasticity." *J. American Statis. Assoc.* 64 (March 1969): 316–23.

Graham, Frank D. *Exchange, Prices and Production in Hyper-Inflation: Germany 1920–1923.* Princeton, N.J.: Princeton Univ. Press, 1930.

Lucas, Robert E., Jr. "Some International Evidence on Output-Inflation Tradeoffs." *A.E.R.* 63 (June 1973): 326–34.

Mills, Frederick C. *The Behavior of Prices.* New York: Arno (for Nat. Bur. Econ. Res.), 1927.

Parks, Richard W. "Inflation and Relative Price Variability." *J.P.E.* 86, no. 1 (February 1978): 79–96.

Phelps, Edmund S. "Introduction: The New Microeconomics in Employment and Inflation Theory." In *Microeconomic Foundations of Employment and Inflation Theory,* by Edmund S. Phelps et al. New York: Norton, 1970.

Sheshinski, Eytan, and Weiss, Yoram. "Inflation and Costs of Price Adjustment." *Rev. Econ. Studies* 44 (June 1977): 287–303.

Statistisches Reichsamt. *Statistisches Jahrbuch für das deutsche Reich.* Berlin: Hobbung, 1921/22 and 1923.

———. *Sonderhefte zur Wirtschaft und Statistic: Zahlen zur Geldenwertung in Deutschland 1914 bis 1923.* Berlin: Hobbung, 1925.

Stigler, George J., and Kindahl, James K. *The Behavior of Industrial Prices.* New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1970.

Vining, Daniel R., and Elwertowski, Thomas C. "The Relationship between Relative Prices and the General Price Level." *A.E.R.* 66 (September 1976): 699–708.

Exchange-Rate Dynamics: An Empirical Investigation

Robert A. Driskill

University of California, Davis

This paper estimates a reduced-form exchange-rate equation whose estimates are used to address questions on exchange-rate overshooting, intermediate-run exchange-rate dynamics, and long-run proportionality relationships between relative money supplies and exchange rates. Based on Swiss–U.S. data from the period 1973–79, the major findings are that following a monetary shock there is short-run exchange-rate overshooting by a factor of about two, and that subsequent exchange-rate adjustments to a new long-run equilibrium take longer than 2 years and exhibit nonmonotonic patterns.

I. Introduction

Much of the recent literature on floating exchange rates has viewed relative money supplies as the basic determinant of exchange rates and relative price levels. While this literature is in agreement about the long-run relationships between money, exchange rates, and price levels, it differs in its predictions about short- and intermediate-run exchange-rate determination.[1] In particular, one strand of this literature, mainly associated with the work of Dornbusch (1976), has emphasized the role of slowly adjusting commodity price levels for short- and intermediate-run exchange-rate dynamics. The most striking implication of the Dornbusch model is that the exchange rate may "overshoot" in the short run: In response to a change in relative

This study stems from my doctoral dissertation. Grateful acknowledgments are due to my primary advisor, Jürg Niehans, and also to Bela Balassa, Jeffrey Frankel, Dale Henderson, George Kanatas, Ed Kane, Jerry Thursby, and an anonymous referee.

[1] For a discussion of these different approaches, see Bilson (1978).

[*Journal of Political Economy*, 1981, vol. 89, no. 2]

money supplies, the exchange rate may immediately change by more than its long-run equilibrium value. Another implication of the model is that, following an initial overshoot, the exchange rate then monotonically approaches its long-run equilibrium value. Both of these implications result from the key assumptions of slowly adjusting commodity prices and perfect capital substitutability and mobility.

Within the theoretical framework of slowly adjusting commodity prices, an alternative to Dornbusch's view of exchange-rate dynamics has developed, an alternative which emphasizes stock/flow interactions and relative-price trade-balance effects. This literature, developed by, among others, Branson (1976), Niehans (1977), and Henderson (1980), alters the strong Dornbusch conclusions about short-run overshooting and monotonic exchange-rate and price-level adjustments to long-run equilibrium, instead permitting short-run undershooting and nonmonotonic exchange-rate and price-level adjustments.

The purpose of this paper is to examine empirically the predictions of these exchange-rate dynamic models against Swiss/U.S. data over the period 1973–77. Specifically, the paper addresses the questions of whether there is short-run exchange-rate overshooting, whether the intermediate-run exchange-rate and price-level adjustments are nonmonotonic, and whether the long-run proportionality relationships between money supplies, price levels, and exchange rates hold.

This is not the first attempt to test a model of exchange-rate determination based on the Dornbusch exchange-rate dynamics approach. Frankel (1979) extends the Dornbusch model by incorporating secular inflation rates and tests it against the deutschmark/dollar rate, and Bilson (1978) provides suggestive evidence about the appropriateness of the Dornbusch approach for the deutschmark/pound rate. Neither test, however, uses an exchange-rate equation rich enough to incorporate trade-balance responses to relative price changes and thus exclude the possibility of more complex, perhaps oscillatory exchange-rate adjustment paths and exchange-rate undershooting.

The major empirical findings from this study are as follows: (1) The exchange rate overshoots in the quarter in which a monetary change takes place by a factor of about two. (2) The exchange-rate adjustment path to full equilibrium is not monotonic, as predicted by the original Dornbusch fixed-output model, but rather exhibits periods of appreciation and depreciation. The price level, however, adjusts monotonically. (3) Purchasing power parity holds in the long run. As noted above, there is wide theoretical agreement that this should be the case, but it has often been difficult to verify empirically. Furthermore, the "long run" is calculated to be approximately 2–3 years.

II. The Reduced-Form Equations

The goal of this section is to motivate the reduced-form exchange-rate and price-level equations which are estimated in the following section. Reduced-form equations are, of course, consistent with a variety of structural models. What is done here is to develop two structural models, the Dornbusch model and a stock/flow model developed by generalizing the Dornbusch model to permit imperfect capital mobility. Both of these structural models impose a priori constraints on the reduced-form parameters and thus can in principle be rejected by the data. Of course, neither structural model may be sufficiently complex, and a "true" structural model may impose few theoretical predictions about a priori reduced-form constraints. This possibility emphasizes the fact that the reduced-form empirical estimates are more general than either structural model.

A. The Dornbusch Model

This subsection recasts the Dornbusch model in discrete time. The model consists of three basic building blocks: a money-market equilibrium condition, a price-level adjustment equation, and an uncovered-interest-arbitrage specification. Assuming that both the domestic and foreign country have identical structural parameters, we can specify money demand as:

$$m_t^d = p_t + \phi y_t - \lambda r_t + v_t, \tag{1}$$

where m is the log of the ratio of the domestic to foreign money supply, p is the log of the ratio of the domestic to foreign price level, y is the log of the ratio of domestic to foreign real income, r is the difference between domestic and foreign interest rates, λ is the interest rate semielasticity of the demand for money in each country, ϕ is the income elasticity of the demand for money in each country, and v_t is a serially uncorrelated random variable with zero mean and variance σ_v^2.

Each country's money supply is assumed to be exogenously controlled by the respective government. Furthermore, all changes in the relative money supply are assumed to be unanticipated; this assumption is important because it implies that this period's relative money supply is also the expected supply for all future periods. Assuming equilibrium obtains each period, we can write the equilibrium condition as:

$$m_t = (m_t)^d = p_t + \phi y_t - \lambda r_t + v_t. \tag{2}$$

In the goods market, relative demand for output depends on relative real income, the relative interest rate, and relative prices and has the following functional form:[2]

$$\log D_t = \gamma y_t - \sigma r_t + \omega(e_t - p_t), \tag{3}$$

where e is the log of the exchange rate and D is demand for domestic output.

The rate of relative inflation, $p_{t+1} - p_t$, is proportional to the log of the ratio of relative demand to relative supply in the goods market:

$$p_{t+1} - p_t = \delta(\log D_t - y_t). \tag{4}$$

Combining equations (2), (3), and (4), we can write the relative price equation as:

$$p_t = a_0 y_{t-1} + a_1 p_{t-1} + a_2 m_{t-1} + a_3 e_{t-1}, \tag{5}$$

where: $a_0 = \delta(1 - \gamma) + \phi/\lambda$, $a_1 = 1 - \delta\sigma/\lambda - \delta\omega$, $a_2 = \delta\sigma/\lambda$, and $a_3 = \delta\omega$.

The final building block of the Dornbusch model is a joint assumption of uncovered interest arbitrage and exchange-rate expectations. The uncovered interest arbitrage assumption can be stated as

$$r_t - x_t = 0, \tag{6}$$

where x_t is the expected change in e from t to $t + 1$. The assumption about x_t is that it is a fraction of the gap between the current and long-run equilibrium exchange rate. Given the assumption that the relative money supply follows a random walk, this implies the following relationship:

$$x_t = \theta(m_t - e_t) + k,\ 0 < \theta < 1,\ k \text{ a constant}. \tag{7}$$

Combining equations (2), (5), (6), and (7), we can derive the following reduced-form equation:

$$e_t = \pi_0 + \pi_1 e_{t-1} + \pi_2 m_t + \pi_3 m_{t-1} + \pi_4 p_{t-1} + \pi_5 y_t + \pi_6 y_{t-1} + \pi_7 z_t, \tag{8}$$

where z_t is a first-order serially correlated random variable and where the π_i's satisfy the following constraints:[3]

[2] Output demand could be specified as a function of the *real* interest rate. In both the Dornbusch and the stock/flow models, this change makes no difference in the qualitative implications derived for the reduced-form equations.

[3] Details are found in Appendix A.

$$\sum_{i=1}^{4} \pi_i = 1 \qquad \pi_3 < 0$$

$$\pi_1 < 0 \qquad \pi_4 < 0$$

$$\pi_2 > 1 \qquad \pi_5 < 0 \qquad \pi_6 < 0.$$

The constraint on the sum of the first four π_i's simply says that purchasing power parity holds in the long run. The constraint that $\pi_2 > 1$ says that there must be short-run overshooting.

B. *The Stock/Flow Model*

As noted in the Introduction, the stock/flow model is developed by generalizing the Dornbusch model to permit imperfect capital mobility. To do this, a net demand for foreign assets is specified as a linear function of the expected net yield:

$$B_t = \eta (x_t - r_t), \qquad \eta > 0. \tag{9}$$

In addition, a trade-balance equation is specified as a linear function of the log of relative prices and the log of relative real incomes:

$$T_t = \alpha (e_t - p_t) - \beta y_t + u_t, \qquad \alpha, \beta \geqslant 0, \tag{10}$$

where u_t is a zero-mean, finite-variance, serially uncorrelated random variable.[4]

The foreign-exchange market-clearing equation states that net capital flows equal net trade flows plus all other autonomous flows (assumed constant):

$$\Delta B_t = T_t + A_t, \tag{11}$$

where A_t is a constant. Replacing equation (6) in the Dornbusch model by equation (11), we can derive the following reduced-form exchange-rate equation:

$$e_t = \pi_0' + \pi_1' e_{t-1} + \pi_2' m_t + \pi_3' m_{t-1} + \pi_4' p_{t-1} + \pi_5' y_t + \pi_6' y_{t-1} + \pi_7' z_t'. \tag{12}$$

The π_i''s satisfy the following constraints:[5]

[4] The term T_t could be specified as a function of current and lagged relative prices where the current-price effect is perverse, i.e., creates a *J*-curve effect. This specification is indistinguishable from (10) in terms of restrictions on reduced-form parameters.

[5] Details are in Appendix A.

$$\sum_{i=1}^{4} \pi_i' = 1 \qquad \pi_3' \gtrless 0$$

$$\pi_1' < 1 \qquad \pi_4' > 0$$

$$\pi_2' > 0 \qquad \pi_5' \gtrless 0 \qquad \pi_6' \gtrless 0.$$

While the purchasing power parity constraint holds just as with the Dornbusch structural model, the other constraints are quite different. In particular, both the lagged exchange-rate and lagged price-level coefficients may be positive in this model. The estimated sign of these coefficients should provide a sharp test between these two models. Also note that π_2' need not be greater than one; this model makes the question of overshooting or undershooting an empirical question.

III. Estimation Results

The reduced-form equations (5), (8), and (12) were estimated for Swiss/U.S. data with two changes. First, after some experimentation with various proxies for Swiss income (data on which are available only on a yearly basis), the income variables y_t and y_{t-1} were dropped. None of the proxies used in the experimentation provided significant coefficients, and the remaining coefficients were hardly affected when income was dropped. Second, two dummy variables were added to the exchange-rate equation: OIL, which takes the value of 1 for December-January-February 1973–74 and zero for all other periods, reflecting the improved attractiveness of dollar-denominated assets following the announcement of the oil embargo and cuts in Arab oil production; and SEAS, which takes the value of 1 every December-January-February and zero all other periods, capturing the pronounced year-end demand for Swiss francs by Swiss firms for end-of-year "window dressing" of their financial statements.[6]

The choice of the Swiss franc/U.S. dollar rate for the period 1973–77 is based on the relative "cleanliness" of the franc float, the substantial variation in the franc-dollar rate, and the independence over this period of the franc and dollar from the "snake."[7]

The estimation uses quarterly average data, where the quarterly data were generated by averaging three monthly end-of-period figures; that is, March-April-May is the average of March, April, and May. This provides a sample size of 19 observations, beginning March

[6] Schiltknecht (1976) discusses this point.

[7] Schiltknecht (1976) indicates that the Swiss have no exchange-rate targets for the Swiss money supply at least for periods longer than a month or two.

1973 and ending November 1977.[8] Even though monthly data would have had the advantage of more observations, the quarterly specification was used since it is more consistent with the theoretical models developed, that is, money-market equilibrium and trade-balance adjustment. While it would be possible to specify partial asset-adjustment schemes and longer trade-balance lags and use the monthly data, this would introduce substantial multicollinearity problems, precluding precise parameter estimates.

Sources and methods of deriving the data are explained in Appendix B. The price index used was the consumer price index, and the monetary aggregate used was M3. Not surprisingly, the results were sensitive to the choice of monetary aggregate: Over the period in question, U.S. M1 demand has not been very stable in comparison with M3 demand.[9] In the presentation of estimated equations, t-statistics are in parentheses beneath the estimated coefficient, and Durbin's H statistic is presented.[10] All estimations were done using White's (1978) SHAZAM regression package.

The estimated relative price equation is:

$$\begin{aligned} p_t = \underset{(1.90)}{.60} &+ \underset{(4.58)}{.76}p_{t-1} + \underset{(1.96)}{.37}m_{t-1} - \underset{(.73)}{.03}e_{t-1}, \\ R^2\,(\text{adj.}) &= .96, \qquad \text{Durbin's } H = 1.67. \end{aligned} \tag{13}$$

Hence, $a_1 = .76$, $a_2 = .37$, and $a_3 = -.03$. Both a_1 and a_2 are significant at the .05 level for a one-tailed test, and the hypothesis that $a_1 + a_2 + a_3 = 1$ is not rejected at the .05 level. Since a_3 in (13) is insignificant and the wrong sign, the equation was also estimated with only p_{t-1} and m_{t-1} as regressors.[11] The estimated equation is:

$$\begin{aligned} p_t = \underset{(1.86)}{.57} &+ \underset{(4.74)}{.78}p_{t-1} + \underset{(1.89)}{.29}m_{t-1}, \\ R^2\,(\text{adj.}) &= .96, \qquad \text{Durbin's } H = 1.65. \end{aligned} \tag{14}$$

Again, both a_1 and a_2 are significant, and the hypothesis that $a_1 + a_2 = 1$ is not rejected at the .05 level. Finally, the equation was estimated

[8] A regression using end-of-period quarterly data from 1973:II through 1977:III is also reported. More will be said about the choice between average and end-of-period at that point.

[9] See Goldfeld (1976) for evidence on this point.

[10] In regressions with a lagged dependent variable, Durbin's H statistic is the appropriate statistic to test serial correlation of the disturbance terms. For a sample as small as used here, however, the usefulness is diminished.

[11] See Appendix C for a derivation of the conditions under which $a_3 \approx 0$, but the trade balance is still a function of relative prices. Basically it requires that the trade balance be a very small fraction of total aggregate demand.

with the constraint that $a_1 + a_2 = 1$ imposed:

$$\underset{(1.84)}{p_t = .57} + \underset{(4.47)}{.71 p_{t-1}} + \underset{(1.86)}{.29 m_{t-1}}. \tag{15}$$

The estimates of the relative price equations, then, are consistent with the model and have good explanatory power.[12]

Preliminary to testing the exchange-rate equation, it is useful to see whether the assumption used in both structural models that the relative money supply follows a random walk, that is, that $m_t = m_{t-1} + x_t$, where x_t is a serially uncorrelated random variable with zero mean, is reasonably consistent with the sample data. To do this, two procedures were used. First, m_t was regressed against a constant term plus the previous lagged values of m_t. A representative result is:

$$m_t = \underset{(.71)}{.15} + \underset{(2.88)}{.99 m_{t-1}} - \underset{(-1.08)}{.54 m_{t-2}} - \underset{(-.14)}{.08 m_{t-3}} + \underset{(1.50)}{.77 m_{t-4}} - \underset{(-.70)}{.29 m_{t-5}} + \underset{(.79)}{.23 m_{t-6}}. \tag{16}$$

Only the coefficient on m_{t-1} is significant, and it is not significantly different from one. The second procedure used a test suggested by Box and Jenkins (1970, p. 291) for randomness in a residual vector from a time-series process. For this test, m_t was regressed on a constant and on m_{t-1} with the following result:

$$m_t = \underset{(.09)}{.01} + \underset{(13.9)}{1.01 m_{t-1}}. \tag{17}$$

Residuals from this equation and from an equation in which the constant term and m_{t-1} coefficient were constrained at (0, 1) were then used in the aforementioned test for nonrandomness. Nonrandomness was rejected for both sets of residuals at the 10 percent level. Consequently, the assumption that all changes in the relative money supply are unanticipated seems justified.

The estimate of the exchange-rate equation is:[13]

[12] The relative price equation can be disaggregated by country and written in the following form: $P_t^{SW} = d_0 + d_1 p_t^{US} + d_2 p_{t-1}^{SW} + d_3 p_{t-1}^{US} + d_4 m_{t-1}^{SW} + d_5 m_{t-1}^{US}$. If the structural equations of Switzerland and the United States are indeed identical, the d_i's should satisfy the following restrictions: $d_1 = 1$, $d_2 = -d_3$, $d_4 = -d_5$. This equation was estimated, and the above hypothesis about the d_i's was not rejected at the .05 level; the standard errors, however, were quite large.

[13] As with the price equation, the exchange-rate equation was also estimated with country-disaggregated data to test the assumption of identical structural equations across countries. The equation is: $e_t = h_0 + h_1 m_t^{SW} + h_2 m_t^{US} + h_3 m_{t-1}^{SW} + h_4 m_{t-1}^{US} + h_5 p_{t-1}^{SW} + h_6 p_{t-1}^{US} + h_7 e_{t-1}$. The coefficient restrictions consistent with identical structural parameters across countries are: $h_1 = -h_2$, $h_3 = -h_4$, $h_5 = -h_6$. The hypothesis test of these restrictions was not rejected at the .05 level; the assumption of identical structural parameters cannot be rejected. Furthermore, the point estimates were rather similar to those of eq. (18) and satisfied all the other a priori coefficient restrictions.

$$
\begin{aligned}
e_t = {} & \underset{(-2.82)}{-2.22} + \underset{(3.65)}{.43e_{t-1}} + \underset{(5.73)}{2.37m_t} - \underset{(5.60)}{2.45m_{t-1}} + \underset{(2.23)}{.93p_{t-1}} \\
& + \underset{(7.61)}{.15OIL} - \underset{(-5.47)}{.06SEAS}; \; RHO = \underset{(1.37)}{.35},
\end{aligned}
\tag{18}
$$

$$R^2 \text{ (adj.)} = .99, \qquad \text{Durbin's } H = .21,$$

where RHO is the estimated first-order autoregressive parameter of the error term.[14]

The overall explanatory power of the equation is quite good. Furthermore, all coefficients except RHO are significantly different from zero at the .05 level for a one-tailed test. The Durbin H statistic indicates no serial correlation in the disturbances.

As far as a priori constraints on the parameters are concerned, first note that the sum of π_1 through π_4 is 1.28, which is insignificantly different from one at the .05 level. That is, the estimated equation is consistent with the purchasing power parity proportionality prediction about changes in relative money supplies and changes in long-run equilibrium exchange rates. To further test this result, equation (17) was also estimated with this constraint imposed:

$$
\begin{aligned}
e_t = {} & \underset{(-2.89)}{-2.38} + \underset{(4.48)}{.55e_{t-1}} + \underset{(4.92)}{2.30m_t} - \underset{(5.79)}{2.69m_{t-1}} + \underset{(1.91)}{.84p_{t-1}} \\
& + \underset{(7.25)}{.16OIL} - \underset{(-5.11)}{.06SEAS}; \; RHO = \underset{(.79)}{.21}.
\end{aligned}
\tag{19}
$$

These coefficients are quite close to the unconstrained ones, thus providing further evidence that the purchasing power parity principle holds.

Both structural models developed in Section II imply that the purchasing power parity constraint should hold. The two models differed sharply, though, on their predictions of the sign of the lagged exchange-rate and lagged price-level coefficients. The estimation results of positive coefficients on these variables are consistent with the stock/flow model but inconsistent with the Dornbusch one. Thus, the data reject the Dornbusch model. Note, though, that the Dornbusch model's major insight about overshooting is verified as an empirical phenomenon.

Two more procedures were carried out to examine further the robustness of the preceding empirical results. First, the exchange-rate equation was estimated using end-of-period rather than average data.

[14] Estimation of (18) was done using a modified Cochrane-Orcutt procedure in the SHAZAM computer program developed by Kenneth White. Grid search and maximum likelihood nonlinear methods produced almost identical estimates.

While using end-of-period data increases the likelihood that any one data point is an outlier, it has the advantage of not possibly confusing anticipated and unanticipated changes in the money supply. The results are:

$$\begin{aligned} e_t = &-3.16 + .56e_{t-1} + 1.69m_t - 2.47m_{t-1} + 1.17p_{t-1} \\ &(-2.66)\,(3.44) \qquad (2.44) \quad (-3.50) \qquad (2.14) \\ &+ .21\text{OIL} - .07\text{SEAS}; \ \text{RHO} = -.14, \\ &(5.63) \quad (-3.44) \qquad\qquad (-.50) \\ &R^2\,(\text{adj.}) = .97, \qquad \text{Durbin's } H = -.34. \end{aligned} \tag{20}$$

The estimated coefficients are significant at the .05 level, have the same signs as those of equation (18), have similar magnitudes, and satisfy the purchasing power parity constraint; indeed, $\Sigma_{i=1}^{4} \pi_i = .94$, which is not significantly different from one at the .05 level.

Second, an in-period dynamic simulation of the exchange-rate equation was performed. This is important because of the presence of a lagged dependent variable and tends to show whether the estimated regression simply picks up the serial correlation in the dependent variable. The forecast error and squared forecast error were then regressed against time to see whether there is a tendency for it to increase with time. The results are as follows:

$$\begin{aligned} \text{Forecast error} &= .05 - .0002(\text{time}), \\ &\quad (3.83)(-.14) \end{aligned} \tag{21}$$

$$\begin{aligned} \text{Forecast error} &= .003 + .00003(\text{time})^2. \\ &\quad (2.15) \quad (.22) \end{aligned} \tag{22}$$

This provides rather strong support for the empirical results.

IV. Concluding Comments

This paper has addressed those issues in exchange-rate determination raised by the recent work on exchange-rate dynamics. The exchange-rate equation developed was flexible enough to be consistent with, among others, the structural model of Dornbusch (1976) and a stock/flow model incorporating features emphasized by Niehans (1977), Henderson (1979), and others. Both structural models implied that a purchasing power parity a priori constraint on the reduced-form parameters should be satisfied; the estimates did not reject this hypothesis. Other a priori constraints implied by the Dornbusch model were rejected, though. There was, however, empirical verification of Dornbusch's overshooting hypothesis.

To help summarize the empirical findings on these questions, estimated coefficients from equations (15) and (18) are used to calculate an exchange-rate and price-level adjustment path following a one-time monetary increase. To do this, equations (15) and (18) are transformed into final form (suppressing all random disturbance terms):

$$e_t = C_0 + \sum_{i=0}^{\infty} \mu_i m_{t-i}, \qquad C_0 \text{ a constant,} \tag{23}$$

$$P_t = C_0' + \sum_{i=1}^{\infty} \epsilon_i m_{t-i}, \qquad C_0' \text{ a constant,} \tag{24}$$

where the μ_i's and ϵ_i's are as follows:

$$\mu_1 = \pi_2,$$

$$\mu_2 = (\pi_3 + \pi_1\pi_2),$$

$$\mu_3 = (\pi_3 + \pi_1\pi_2)\pi_1 + a_2\pi_4,$$

$$\mu_4 = (\pi_3 + \pi_1\pi_2)\pi_1^2 + a_2\pi_4(\pi_1 + a_1),$$

$$\mu_5 = (\pi_3 + \pi_1\pi_2)\pi_1^3 + a_2\pi_4(\pi_1^2 + a_1\pi_1 + a_1^2),$$

$$\vdots$$

$$\mu_n = (\pi_3 + \pi_1\pi_2)\pi_1^{n-1} + a_2\pi_4(\pi_1^{n-2} + \pi_3^{n-3}a_1 + \pi_1^{n-4}a_1^2 + \ldots + \pi_1 a_1^{n-3} + a_1^{n-2}),$$

$$\epsilon_1 = a_2,$$

$$\epsilon_2 = a_2 a_1,$$

$$\epsilon_3 = a_2 a_1^2,$$

$$\vdots$$

$$\epsilon_n = a_2 a_1^{n-1}.$$

Using these final form coefficients and assuming the initial exchange rate and relative price level are ($e = 0$, $p = 0$), then, following a one-unit unanticipated monetary increase, we can calculate the adjustment path as follows:

$$
\begin{array}{lll}
e_0 = 2.30 & p_0 = 0 & e_0 - p_0 = 2.30 \\
e_1 = .88 & p_1 = .29 & e_1 - p_1 = .59 \\
e_2 = .34 & p_2 = .50 & e_2 - p_2 = -.16 \\
e_3 = .21 & p_3 = .65 & e_3 - p_3 = -.44 \\
e_4 = .27 & p_4 = .75 & e_4 - p_4 = -.48 \\
e_5 = .39 & p_5 = .82 & e_5 - p_5 = -.43 \\
e_6 = .48 & p_6 = .87 & e_6 - p_6 = -.39 \\
e_7 = .54 & p_7 = .91 & e_7 - p_7 = -.37 \\
e_8 = .61 & p_8 = .94 & e_8 - p_8 = -.33 \\
e_9 = .67 & p_9 = .96 & e_9 - p_9 = -.29 \\
e_{10} = .73 & p_{10} = .97 & e_{10} - p_{10} = -.24.
\end{array}
$$

The exchange rate initially depreciates to 2.30, then appreciates for three quarters to 0.21, and then depreciates again. After 11 quarters, about three-quarters of the exchange-rate adjustment has taken place, and 97 percent of the price adjustment has occurred. Notice that deviations from purchasing power parity are positive for only the first two quarters, after that remaining negative. This means that exports are stimulated and imports suppressed because of increased relative prices for only one quarter, and then the effects are reversed.[15] Note also that the exchange-rate path to long-run equilibrium is, as anticipated, nonmonotonic, showing periods of both appreciation and depreciation.

To keep these empirical results in perspective, some of the more serious limitations should be noted. First, the exchange-rate equation tested was based on a bilateral model, while in reality the $n - 1$ independent exchange rates of an n-country world are functions of exogenous variables of all n countries. The equations estimated here, then, are subject to omitted-variables specification bias. If both the correlation of these omitted variables with the included variables and the expected sign of the coefficient of these omitted variables can be determined, then the bias in the estimated coefficients can be determined. At present, not even the expected sign of the omitted coefficients has been determined.[16]

Second, the equation estimated is a reduced form from whose

[15] Note that the cumulative trade balance over the adjustment path must be zero, since there is no net accumulation of foreign assets.

[16] See Berner et al. (1975) for an attempt to develop an n-country model.

estimates structural parameter values cannot be retrieved. While the reduced-form estimates seem plausible, they would be more credible if they were shown to be consistent with existing estimates of an underlying structural model's parameters. Still, the results provide general support for the monetary approach to exchange-rate determination and specific support for the importance of both slow commodity-price adjustments and current-account relative-price effects within that approach.[17]

Appendix A

I. *Derivation of Reduced-Form Constraints from the Dornbusch Model*

Combining equation (7) with (6) and (2), we get:

$$e_t = m_t(1 + 1/\theta\lambda) + p_t(-1/\theta\lambda) + y_t(-\phi/\theta\lambda). \tag{A1}$$

Substituting (5) for p_t we have:

$$\begin{aligned} e_t = {} & e_{t-1}(-a_3/\theta\lambda) + m_t(1 + 1/\theta\lambda) + m_{t-1}(-a_2/\theta\lambda) + p_{t-1}(-a_1/\theta\lambda) \\ & + y_t(-\phi/\theta\lambda) + y_{t-1}(-a_0/\theta\lambda). \end{aligned} \tag{A2}$$

II. *Derivation of Reduced-Form Constraints from the Stock/Flow Model*

Replacing equation (6) of the Dornbusch model with equation (11) we get the following difference equation:

$$\begin{aligned} e_t = {} & e_{t-1}\left[\frac{\eta\theta + a_3(\alpha - \eta/\lambda)}{\eta\theta + \alpha}\right] + m_t\left(\frac{\eta\theta + \eta/\lambda}{\eta\theta + \alpha}\right) \\ & + m_{t-1}\left(\frac{-\eta\theta - \eta/\lambda - a_2\eta/\lambda + a_2\alpha}{\eta\theta + \alpha}\right) + p_{t-1}\left[\frac{(1 - a_1)\eta/\lambda + \alpha a_1}{\eta\theta + \alpha}\right] \\ & + y_t\left(\frac{\beta}{\eta\theta + \alpha}\right) + y_{t-1}\left[\frac{a_0(\alpha - \eta/\lambda)}{\eta\theta + \alpha}\right]. \end{aligned} \tag{A3}$$

Note that the coefficient on m_t may be less than one, the coefficient of e_t may be positive, and the coefficient on p_{t-1} must be positive.

Appendix B

I. *Switzerland*

1. *Exchange rate.*—End-of-period data are end-of-month single observations. Source: OECD Main Economic Indicators, Paris, various issues. Quarterly data are averages of daily noon New York rates. Source: *Federal Reserve Bulletin,* Washington, various issues.

2. *Money supply.*—The M3 is taken from International Financial Statistics of the IMF. It appears to include some demand deposits denominated in

[17] Again, see Bilson (1978) for a discussion of how the Dornbusch approach fits within the general monetary framework.

foreign currencies. From figures supplied by Peter Buomberger of the Swiss National Bank, it appears that this component is a small percentage of total M3 (less than 4 percent) and fairly stable. These figures were only for 22 nonconsecutive months, which precluded using them for estimation purposes.

3. *Consumer price index.*—Source: OECD Main Economic Indicators. These are monthly end-of-period figures.

II. *United States*

1. *Money supply.*—Source: *Federal Reserve Bulletin.* Seasonally unadjusted data were used for the quarterly estimations.

2. *Consumer price index.*—Source: OECD Main Economic Indicators.

Appendix C

This Appendix shows that a_3 in the price equation may be close to zero and the trade balance still a function of relative prices if the trade balance is a relatively small fraction of aggregate demand for a country's output. Let D be the ratio of domestic to foreign aggregate demand, E and E^* domestic and foreign expenditure, respectively, that is, $C + I + G$, and P relative prices between the two countries' outputs. The question is what implicit assumption makes $(d \log D)/(d \log P)$ close to zero. Now,

$$D = \frac{E + T}{E^* - T}, \tag{C1}$$

where T is the trade balance. Hence,

$$\begin{aligned} \frac{d \log D}{d \log P} &= \frac{1}{(E + T)} \frac{d(E + T)}{d \log P} - \frac{1 d(E^* - T)}{(E^* - T) d \ln P} \\ &= \frac{P}{(E + T)} \frac{d(E + T)}{dP} - \frac{P}{(E^* - T)} \frac{D(E^* - T)}{dP}. \end{aligned} \tag{C2}$$

Now define the trade balance elasticity as $PT'/T = \epsilon$. Assuming that $dE/dP = dE^*/dP = 0$ and that $E = E^*$, we have

$$\frac{d \log D}{d \log P} = \frac{2\epsilon}{(E/T) - (T/E)}. \tag{C3}$$

Hence, $(d \log D)/(d \log P)$ goes to zero as E/T goes to infinity, despite the size of ϵ. Note that this does *not* say that trade balance responsiveness to relative prices does not affect exchange rates; it only does not affect aggregate demand. What is more important for exchange-rate determination are the relative sizes of trade flows and capital flows.

References

Berner, Richard B.; Clark, P.; Howe, Howard; Kwack, Sung Y.; and Stevens, G. "Simultaneous Determination of the U.S. Balance of Payments and Exchange Rates: An Exploratory Report." International Finance Discussion Paper no. 59, Board of Governors of the Federal Reserve, Washington, February 1975.

Bilson, John F. O. "The Current Experience with Floating Exchange Rates: An Appraisal of the Monetary Approach." *A.E.R. Papers and Proc.* 68 (May 1978): 392–97.

Box, George E. P., and Jenkins, Gwilym M. *Time Series Analysis, Forecasting and Control.* San Francisco: Holden-Day, 1970.

Branson, William H. "Asset Markets and Relative Prices in Exchange Rate Determination." Seminar Paper no. 66, Institute for International Economic Studies, Stockholm, 1976.

Dornbusch, Rudiger. "Expectations and Exchange Rate Dynamics." *J.P.E.* 84, no. 6 (December 1976): 1161–76.

Frankel, Jeffrey A. "On the Mark: A Theory of Floating Exchange Rates Based on Real Interest Differentials." *A.E.R.* 69 (September 1979): 610–22.

Goldfeld, Stephen M. "The Case of the Missing Money." *Brookings Papers Econ. Activity*, no. 3 (1976), pp. 683–730.

Henderson, Dale. "The Dynamic Effects of Exchange Market Intervention Policy: Two Extreme Views and a Synthesis." In *Economics of Flexible Exchange Rates*, edited by Helmut Frisch and Gerhard Schwödiauer (*Kredit und Kapital*, vol. 6 [suppl.], 1980).

Niehans, Jürg. "Exchange Rate Dynamics with Stock/Flow Interaction." *J.P.E.* 85, no. 6 (December 1977): 1245–57.

Schiltknecht, K. "Monetary Policy under Flexible Exchange Rates: The Swiss Case." Mimeographed. Swiss National Bank, 1976.

White, Kenneth J. "A General Computer Program for Econometric Methods—SHAZAM." *Econometrica* 46 (January 1978): 239–40.

The Family as an Incomplete Annuities Market

Laurence J. Kotlikoff

Yale University and National Bureau of Economic Research

Avia Spivak

Ben Gurion University

Families can self-insure against uncertain dates of death through implicit or explicit agreements with respect to consumption and interfamily transfers. Interfamily transfers need have nothing to do with altruistic feelings; they may simply reflect risk-sharing behavior of completely selfish family members. Although family annuity markets are incomplete, even small families can substitute by more than 70 percent for perfect market annuities. Given adverse selection and transaction costs, family risk pooling may be preferred to public market annuities. In the absence of public annuities, these risk-sharing arrangements provide powerful incentives for marriage and family formation.

The institution of the family provides individuals with risk-sharing opportunities which may not otherwise be available. Within the family there is a degree of trust and a level of information which alleviates three key problems in the provision of insurance by markets open to the general public, namely, moral hazard, adverse selection, and deception. In addition, provision of insurance within the family may entail smaller transaction costs than arise in the purchase of insurance on the open market. There are a number of important risks for which

We are grateful for financial support from the Foundation for Research in Economic Education and the National Bureau of Economic Research. Any opinions expressed are solely our own. We wish to thank Finis Welch, Joe Ostroy, Bryan Ellickson, John McCall, Steven Shavell, John Riley, Jon Skinner, and Gary Galles for helpful discussions.

[*Journal of Political Economy*, 1981, vol. 89, no. 2]

the "public" market problems of moral hazard, adverse selection, and deception are especially severe. The risk of loss of job or earnings because of changes in the pattern of demand or partial disability is one example. Here the ability of the public market to determine the extent to which the individual actually suffered an earnings loss or is simply lying about his backache is highly questionable. Other examples are the risk of bankruptcy and the default risk on personal loans. Many family practices in dealing with these types of risks can be explained as implicit insurance contracts made ex ante by completely selfish family members. Love and affection may be important for the enforcement of some of these implicit contracts, but they need not be their sole or even chief determinant. Healthy brother A's support for disabled brother B may simply be the quid pro quo for brother B's past implicit promise to support A if A became disabled instead of B.

The existing economics literature on marriage and the family (including Schultz [1974] and Becker, Landes, and Michael [1977]) has not, to our knowledge, explicitly considered the family's role in providing insurance to family members.

This paper is concerned with family provision of insurance against the risk of running out of consumption resources because of greater than average longevity. The problem is how fast to consume over time when one does not know how long one will continue to live. Too much consumption when young may mean relative poverty later on if one lives "too long"; alternatively, excessive frugality when young involves the risk of dying without ever having satisfied one's hunger. A complete annuity market permits an individual to hedge this uncertainty of the date of death by exchanging his initial resources for a stream of payments that continue as long as the individual survives. We demonstrate here that implicit risk-sharing arrangements within marriage and the family can substitute to a large extent for the purchase of annuities in public markets. Since the number of family members involved in the risk pooling is generally small, these family risk-sharing arrangements constitute an incomplete annuities market. However, our findings suggest that even small families can substitute by more than 70 percent for a complete annuity market in pooling the risk of death. When the economic structure of society is sufficiently developed to sustain organized public insurance markets, implicit risk pooling within an incomplete family annuity market may well be preferred to public purchase of annuities because of adverse selection and transaction costs.[1] When organized insurance markets do not

[1] The transaction costs we have in mind here include the time costs involved in negotiating individual specific annuity contracts. As we demonstrate in the text, each individual's optimal annuity contract depends on his rate of time preference, his degree

exist, the analysis here indicates that implicit risk-sharing arrangements can provide powerful economic incentives for marriage and family formation.

Throughout the paper individuals are assumed to be completely selfish; that is, they obtain utility only from their own consumption. One implication of this approach is that voluntary transfers from children to parents or bequests and gifts from parents to children need have nothing at all to do with altruistic feelings; rather, they may simply reflect risk-sharing behavior of completely selfish individuals. While altruism per se is not required, some level of mutual trust and honesty *is* required since elements of these arrangements are not legally enforceable.

This paper is divided into four sections, the first of which describes optimal consumption behavior for a single individual in both the presence and absence of a complete annuity market. The welfare gains from access to a complete and fair annuity market are calculated for the case of the iso-elastic utility function. This welfare gain is, in turn, decomposed into income and substitution effects. This decomposition suggests that an important component of the gains from access to complete or incomplete annuity markets is the desirability of substituting future for current consumption.

Section II develops the theoretical argument for Pareto-efficient implicit family annuity contracts and explores potential welfare gains arising from these arrangements.[2] Although the complexity of the calculations precluded analysis of large families, quantitative results for families of two and three persons are presented. The analysis considers cases in which family members both do and do not have identical survival probabilities (i.e., are of similar and dissimilar ages and sexes). This framework permits us to ask whether marriage between individuals with similar survival probabilities is more efficient than marriage between individuals with dissimilar survival probabilities.

Optimal family annuity contracts involve agreements on the consumption path of each family member as well as a commitment on the part of each member to name the other members as sole heirs in his estate. Section III discusses the problems of enforcing both aspects of these agreements. Section IV summarizes the paper and suggests areas for future research.

---

of risk aversion, and his survival probabilities. Some individuals may prefer a constant annuity stream, others an increasing or decreasing stream of annuity payments.

[2] Kotlikoff and Spivak (1979) present a proof that family annuity contracting converges to a complete annuities market as the number of family members increases.

I. A Single Person's Consumption Plans with and without Fair Annuities[3]

In the absence of an annuity market, a single individual's consumption choice problem is to maximize his expected utility, equation (1), from current and future consumption subject to the budget constraint, equation (2):

$$EU = \sum_{t=0}^{D} P_t U(C_t), \tag{1}$$

$$\sum_{t=0}^{D} C_t R^{-t} = W_0. \tag{2}$$

The P_t's of equation (1) are probabilities of surviving from age zero through age t; P_0 equals one. The term D is the maximum longevity. For simplicity, we assume the utility function is separable in consumption (C_t) over time. In (2) R, the discount factor, is one plus the interest rate. The initial wealth of the individual is W_0; we ignore possible streams of future labor earnings or inheritances.[4]

The budget constraint written in equation (2) is identical to the budget constraint that would arise in a certainty world in which individuals never died before age D. While individuals will, on the average, die prior to age D, equation (2) reflects the nonzero probability that an individual will live through age D; that is, equation (2) is the relevant budget constraint because the individual may actually live through age D, in which case his realized present value of consumption cannot exceed his budget.

Let us now assume that the single person is free to purchase actuarially fair annuities in a complete public annuities market. The budget constraint in this case is

$$\sum_{t=0}^{D} P_t C_t R^{-t} = W_0. \tag{3}$$

[3] Yaari (1965) is the pioneering paper on this subject. Sheshinski and Weiss (1981) provide an illuminating discussion on the interaction of annuities and social insurance. Barro and Friedman (1977) provide an analysis of the risks of the uncertainty of the date of death.

[4] The gains from access to an annuities market are greatest when the individual has all his resources up front. This assumption, then, dramatizes the demand for annuities; but dropping this assumption would not alter the theoretical point that families can substitute for annuity markets. For the sake of completeness one can think of the individuals described in this paper as having received all their resource streams prior to their current age. In the no-annuity, no-family world involuntary bequests can be thought of as being collected by the government and redistributed to individuals at their birth.

In contrast with (2), (3) requires only an equality between the expected present value of consumption and initial wealth. The single individual now chooses his optimal consumption path by maximizing (1) subject to (3); he then exchanges his initial wealth W_0 with the insurance company in return for its promise to pay out the C_t stream as long as the person continues to live.

The P_tR^{-t}'s in (3) may be thought of as prices. Since each of the P_t's in (3), except P_0 which equals unity, is less than one, the consumption choice in the case of a fair annuity market is equivalent to the consumption choice without an annuity market but with lower prices of future consumption. Obviously, access to a fair annuity market increases utility by expanding the budget frontier; it also alters the optimal consumption path because of the income and substitution effects resulting from the lower prices of future consumption.

The iso-elastic utility function (4) is convenient for assessing the potential gains from access to a fair public annuities market as well as the gains from family annuity arrangements:

$$EU = \sum_{t=0}^{D} P_t \frac{C_t^{1-\gamma}}{1-\gamma} \alpha^t. \tag{4}$$

In (4), γ is the constant relative risk-aversion parameter, and α is the time preference parameter. By considering different values of γ we indicate for this family of utility functions how the gains from annuities and family arrangements depend on the specification of tastes.

In the no-annuities case maximization of (4) subject to (2) leads to the consumption plan, (5):

$$C_t = \frac{W_0(R\alpha)^{t/\gamma}P_t^{1/\gamma}}{\sum_{j=0}^{D} R^{j(1-\gamma)/\gamma}\alpha^{j/\gamma}P_j^{1/\gamma}}. \tag{5}$$

In the case of fair annuities, maximizing (4) subject to (3) leads to

$$C_t = \frac{W_0(R\alpha)^{t/\gamma}}{\sum_{j=0}^{D} R^{j(1-\gamma)/\gamma}\alpha^{j/\gamma}P_j}. \tag{6}$$

Figure 1 compares equations (5) and (6) for the case $R = \alpha = 1$. The ability to trade in a fair annuities market may raise or lower initial consumption, depending on whether γ is less than or greater than unity (fig. 1). Intuitively, the higher the degree of risk aversion, γ, the greater the concern for running out of money because of excessive longevity and, hence, the lower the initial consumption. At γ equal to infinity, equation (5) dictates equal consumption in each period.

Plugging (5) or (6) into (4), we arrive at two indirect utility functions for the no-annuity and annuity cases with initial wealth, the interest

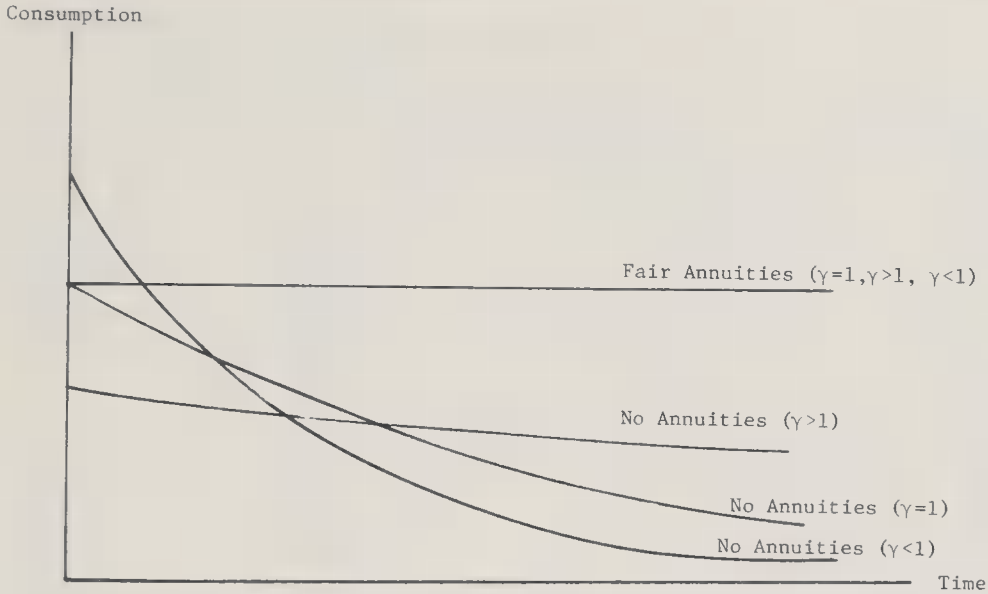


FIG. 1.—Consumption paths with and without fair annuities

rate, and survival probabilities as arguments. These functions are presented in equations (7) and (8), respectively,

$$H_0(W_0) = \frac{1}{1-\gamma} W_0^{1-\gamma} \left[\sum_{j=0}^{D} \alpha^{j/\gamma} R^{j(1-\gamma)/\gamma} P_j^{1/\gamma} \right]^{\gamma}, \tag{7}$$

$$V_0^*(W_0) = \frac{1}{1-\gamma} W_0^{1-\gamma} \left[\sum_{j=0}^{D} \alpha^{j/\gamma} R^{j(1-\gamma)/\gamma} P_j \right]^{\gamma}. \tag{8}$$

The increase in utility resulting from access to fair annuities can be measured in terms of dollars. Equation (9) solves for the value of M, which represents the percentage increment in a single person's initial wealth required, in the absence of an annuity market, to leave him as well off as he would be with no additional wealth but with access to an annuities market:

$$H_0(MW_0) = V_0^*(W_0). \tag{9}$$

For the iso-elastic utility function this calculation is independent of the initial level of wealth. Table 1 reports values of M for different ages and levels of risk aversion using both male and female survival probabilities. Friend and Blume (1975) estimate the degree of relative risk aversion from individual portfolio choices. They conclude that risk aversion, on average, exceeds unity. We present our results for risk-aversion coefficients of 0.75, 1.25, and 1.75, a range that we feel

TABLE 1

Percentage Increase in Initial Wealth Required to Obtain Fair Annuities Utility Level

Age	Relative Risk Aversion (γ)	Males	Females
30	.75	24.5	18.5
55	.75	46.9	34.4
75	.75	71.2	63.0
90	.75	99.8	100.2
30	1.25	30.3	22.7
55	1.25	59.2	43.4
75	1.25	97.0	85.3
90	1.25	152.6	152.9
30	1.75	34.7	26.1
55	1.75	68.9	50.7
75	1.75	119.1	104.6
90	1.75	199.1	199.4

Note.—Throughout table α = .99 and R = 1.01.

encompasses reality. The survival probabilities used in this and all subsequent calculations are actuarial estimates from the Social Security Administration.[5] Maximum longevity is taken to be 120 throughout the paper.

Table 1 indicates that the utility gain measured in dollars from access to an annuities market can be quite large. For a relative risk-aversion parameter value of 0.75, the gain to a 55-year-old male is equivalent to a 46.90 percent increase in his initial wealth. The utility gain is age dependent; for γ = 0.75, the 30-year-old male's gain is 24.46 percent, while the 90-year-old male's gain is 99.81 percent. Annuities are less important to young people because a large fraction of their lifetime utility from consumption is fairly certain due to their lower mortality probabilities in the immediate future. Higher levels of risk aversion naturally increase the gains from access to an annuities market. The male-female differences in the table reflect the higher male age-specific mortality rates. The calculation is somewhat sensitive to the choice of α and R. Raising the interest rate to 5 percent while holding α constant increases the age 55 wealth-equivalent factor from 46.90 to 55.57 for the case of γ = 0.75. The 90-year-old wealth equivalent is increased from 99.81 to 115.34.

[5] We use the low mortality male and female probabilities reported on pp. 17 and 19 of the Social Security Administration Actuarial Study no. 62 (see U.S. Department of Health, Education, and Welfare 1966).

Income, Substitution Effects, and Unintended Bequests

Without access to an annuity market a single, nonaltruistic individual will always die prior to consuming all his wealth and, accordingly, will make involuntary bequests. The level of these unintended bequests can be quite large. From equation (5) we calculated the consumption path as well as the corresponding wealth path for the no-annuity case. By multiplying the probability of dying at each age times the wealth at each age and discounting back to the initial age, the present expected value of these unintended bequests can be computed. For $\gamma = 0.75$, $R = 1.01$, and $\alpha = .99$, the present expected value of unintended bequests represents 24.47 percent of initial wealth for a single male aged 55. This number means that a 55-year-old male with no annuity market will, on average, fail to consume about one-quarter of his wealth because he is risk averse. Increasing the risk-aversion coefficient to 1.75 raises the ratio of present unintended bequests to initial wealth to 0.3583. These large unintended bequests occur despite a fairly rapid rate of consumption. Current mortality probabilities dictate a fairly rapid rate of consumption even for high levels of risk aversion. For $\gamma = 1.75$, a single male who survives to age 85 consumes at age 85 less than a third of his age 55 consumption level.[6]

The homothetic property of the iso-elastic utility function permits a decomposition of the utility gains from fair annuities into income and substitution effects. Suppose a fair insurance company approached a single, 55-year-old ($\gamma = 0.75$) male and offered to pay him 24.47 percent of his initial wealth in exchange for his naming the insurance company as his heir. The single male would take the 24.47 percent gain and, because of homotheticity, consume it according to his original no-annuity consumption path. This additional wealth would give rise to an additional $.2447 \times .2447$ in present expected bequests. By letting the insurance company also pay for this second round of expected but involuntary bequests as well as further rounds, the insurance company ends up paying $32.40 = .2447/(1 - .2447)$ percent of the single individual's initial wealth. This 32.40 percent figure represents the utility gain from the pure income effect. In this scenario the individual continues to consume at the no-annuity set of prices. Since the total gain from being able to purchase fair annuities and thus face lower prices for future consumption is 46.90 percent, the income effect represents 69.08 percent and the substitution effect 30.92 percent of the total gain. Hence, the ability to alter the age

[6] The ratio of consumption at age 75 to consumption at age 55 is 0.62. When risk aversion equals 0.75, the ratio of consumption at age 85 to consumption at age 55 equals 0.06; it is 0.33 at age 75.

consumption profile is an important part of the total welfare gain from annuities.

II. The Family as an Incomplete Annuities Market

Decisions by family members concerning consumption expenditures and interfamily transfers may reflect implicit though incomplete annuity contracts. In the case of marriage both individuals commonly agree to pool their resources while both marriage partners are alive and to name each other as the major, if not the sole, beneficiary in their wills. For each partner the risk of living too long is somewhat hedged by the other partner's potential death; if one partner lives to be very old, there is a high probability that his (or her) spouse has already died leaving him a bequest to help finance his consumption. While each spouse gains simply from the exchange of wills, the two can further increase their expected utilities by agreeing on a joint consumption path that takes into account each spouse's expected bequest to the other. The importance of joint consumption planning is highlighted in the case of an implicit contract between a parent and a child. Here the parent implicitly promises to name the child in his will in exchange for the child's implicit promise to care for the parent if the parent lives too long. Although the child may have zero probability of dying while the parent is still alive, both can gain because the child agrees to share consumption resources with the parent.

This view of bequest and consumption arrangements within marriage as an incomplete annuity market becomes intuitive when one contemplates increasing the number of members in the family. To simplify the issue, let us assume that all individuals within the family have identical survival probabilities and that they enter this multiperson family with identical resources. In the limit as the family (or "tribe") gets large, the consumption path of an individual within the tribe converges to the path a single individual would choose in a complete and actuarially fair annuities market (Kotlikoff and Spivak 1979).

Quantitative Analysis of Family Risk Pooling

In the case of two family members the frontier of efficient marriage contracts is obtained as the solution to the following recursive dynamic programming problem:

$$V_{t-1}(W_{t-1}) = \max_{W_t, C^H_{t-1}, C^S_{t-1} \geq 0,\, t=T,\ldots,1} [u^H(C^H_{t-1}) + \theta u^S(C^S_{t-1}) + \alpha P_{t/t-1} Q_{t/t-1} V_t(W_t) + \alpha P_{t/t-1}(1 - Q_{t/t-1}) H_t(W_t) + \theta \alpha Q_{t/t-1}(1 - P_{t/t-1}) S_t(W_t)], \tag{10}$$

subject to

$$W_t/R + C^H_{t-1} + C^S_{t-1} = W_{t-1}, \tag{11}$$

where

$$V_T(W_t) = \max_{C^H_T, C^S_T} u^H(C^H_T) + \theta u^S(C^S_T).$$

In (10) $V_t(W_t)$ is the period t maximum-weighted expected utility of the two family members with joint wealth W_t. In the expression the letters H and S denote the two family members, C^H_t and C^S_t are the consumptions of the two, u^H and u^S are their utility functions, $P_{t/t-1}$ and $Q_{t/t-1}$ are their respective period t survival probabilities conditional upon surviving through period $t - 1$, and $H_t(W_t)$ and $S_t(W_t)$ are the maximum expected utilities for each member if he or she alone survives to period t. These expressions are obtained from equation (7) by replacing W_0 with W_t and applying the appropriate probabilities. The term θ is the differential weight applied to member S's expected utility.

The first two terms on the right-hand side of (10) represent utility from certain period $t - 1$ consumption. The third term is the family's expected period t utility multiplied by the probability that both members survive to period t. The last two terms represent expected utilities when one member dies and the other survives.

The Appendix presents an algorithm to solve (10). The algorithm for solving the three-family-member maximization problem is available from the authors.[7]

The Gains from Family Annuity Contracts

The solution to (10) permits a comparison of consumption paths and utility levels of married people with those of single persons, assuming throughout that there is no public annuities market. Both spouses are assumed to have identical iso-elastic utility functions in the sense of the same degrees of risk aversion and rates of time preference.

The shape of consumption paths for married couples while they are both alive may differ from that of single individuals for two reasons. First, even if the two spouses have identical survival probabilities, the reduction in risk within the marriage rate acts like a reduction in the price of future consumption. If the relative risk-aversion parameter γ exceeds (is less than) one, the identical survival probabilities marriage profile will start above (below) the single person's profile. For γ equal

[7] In the three-member family we maximize a weighted sum of the three members' expected utility taking all survival contingencies into account. If one of the three dies first, the other two jointly inherit the remaining wealth and consume according to the optimal two-person plan.

to unity the profiles are identical. In terms of figure 1, the consumption profiles for married persons lie between the no-annuity and complete annuity profiles. The second reason for different consumption profiles for married people relative to single individuals is possible differences in spousal survival probabilities. Higher survival probabilities act like lower rates of time preferences. When an old man marries a young woman the slope of the optimal marriage consumption profile reflects the survival probabilities of both the old husband and the young wife. The two spouses compromise with respect to the rate at which they eat up their joint wealth while they are both alive. The old husband would prefer to eat up the wealth more rapidly, and the young wife would prefer to consume at a slower rate. The formula for each spouse's consumption when married takes both spouses' survival probabilities into account as well as the relative spousal utility weights. To our knowledge empirical studies of consumption and savings at the household level have not considered this point—that the time preference rate for a household may depend on the age-sex composition of the household.

Table 2 reports the gains from marriage as well as three-person polygamy among individuals who have identical survival probabilities and identical initial endowments and who are weighted equally in the contract. The marriage and three-person polygamy gains are calculated as the percentage increase in a single person's initial wealth needed to make him as well off as he would be in the marriage or polygamous relationship. The table also reports the dollar gain as a fraction of the table 1 total dollar gain from complete and fair annuities. Since utility is concave in wealth, the dollar gain from these family contracts as a fraction of the dollar gain from fair annuities is smaller than the actual utility gain from these contracts as a fraction of the utility gain from fair annuities. Table 2 also reports this latter fraction.[8]

The figures in table 2 indicate that marriage can offer substantial risk-pooling opportunities. For a 55-year-old man using male survival probabilities, pooling risk through marriage is equivalent to about a 20 percent increase in his wealth had he stayed single. The gains from marriage increase as one becomes older since the risks incurred are much greater as one ages. At age 75 marriage is equivalent to increasing one's wealth by 30 percent when risk aversion is 1.25. Death-risk-pooling through marriage can be quite important even at young ages. The table reports gains from 11.7 to 13.6 percent at age 30 using the male probabilities.

[8] This fraction is calculated as $[(1 + m)^{1-\gamma} - 1]/[(1 + a)^{1-\gamma} - 1]$, where m is the fractional wealth equivalent gain from marriage, and a is the fractional wealth equivalent gain from fair annuities.

TABLE 2

THE ANNUITY GAINS FROM MARRIAGE AND THREE-PERSON POLYGAMY

		Marriage			Three-Person Polygamy		
Age	Risk Aversion	Dollar Marriage Gain (%)	Dollar Marriage Gain/Dollar Annuity Gain	Utility Marriage Gain/Utility Annuity Gain	Dollar Polygamy Gain (%)	Dollar Polygamy Gain/Dollar Annuity Gain	Utility Polygamy Gain/Utility Annuity Gain
30	.75	11.7	.478	.499	15.8	.645	.665
55	.75	20.0	.426	.461	28.0	.597	.632
75	.75	25.4	.357	.405	37.2	.522	.571
90	.75	28.2	.283	.339	43.1	.432	.496
30	1.25	13.0	.429	.470	18.0	.594	.635
55	1.25	22.3	.377	.446	32.1	.542	.612
75	1.25	30.1	.310	.419	45.2	.466	.571
90	1.25	37.1	.243	.367	58.2	.381	.524
30	1.75	13.6	.392	.456	19.3	.556	.619
55	1.75	23.5	.341	.451	34.5	.501	.613
75	1.75	33.2	.279	.449	50.7	.426	.596
90	1.75	43.0	.216	.420	68.7	.345	.579

NOTE.—The table uses male mortality probabilities. $R = 1.01$ and $\alpha = .99$.

Marriage can also close much of the utility gap between no annuities and complete annuities. For example, for a 55-year-old with risk aversion of 0.75, marriage substitutes 46.10 percent for complete and fair annuities. Marriage is a better substitute for fair annuities at younger ages because at younger ages the probability that both spouses will die simultaneously is quite small relative to the probability that one spouse will die before the other. In addition, there appears to be an interaction between age and the degree of risk aversion, making marriage a better substitute for fair annuities at young ages when risk aversion is low and at old ages when risk aversion is high.

Over a wide range of ages and parameter values, three people appear to be capable of capturing about 60 percent of the gains from fair annuities. While the complexity of the calculations precluded considering a four-person arrangement, we can conjecture using table 2 how well four people would do together. In the case of a 55-year-old male with risk aversion of 0.75, adding one marriage partner is equivalent to a 20 percent increase in his wealth had he remained single. The marginal dollar gain from adding a third person (table 2) is 8.04 percent. If the marginal dollar gain fell at a constant rate in this range, the fourth person would add $8.04 \times (8.04/20.0) = 3.23$ percent.[9] By adding 3.23 to 28.04, we can roughly calculate the extent to which four people can close the utility gap. The procedure suggests that four people can substitute by 70 percent for a fair annuities market.

Diminishing returns to risk pooling appear, then, to set in at a fairly rapid rate. In this example two people substitute by 46 percent, three people by 63 percent, and four people by over 70 percent for full insurance.

Table 3 considers incomplete annuity arrangements between two parents and one child and between one parent and two children. In both cases we assume equal consumption by all family members but permit the initial wealth of the child or children to vary. All individuals are assigned the male survival probabilities; the children are age 30 and the parents age 55. In the case of two parents with one child, if the child has an initial wealth of $35,000 and the parents have an initial wealth of $20,000, entering into an equal consumption–will-swapping arrangement is equivalent to a 32 percent increase in wealth for each parent and a 10.6 percent increase for each child. For the parent this arrangement captures 71.2 percent of the utility gain from

[9] This is probably a lower bound estimate for the contribution of the fourth person; the marginal dollar gain cannot fall at a constant 40 percent rate forever, because if it did the total dollar gains would, in the limit, not sum up to 46.9 percent, the full annuity gain of table 1. Presumably the marginal dollar gain falls at a decreasing rate, and 3.23 percent probably underestimates the fourth person's marginal contribution.

TABLE 3

GAINS FROM INCOMPLETE ANNUITY ARRANGEMENTS IN THE FAMILY

Initial Wealth of Each Child ($)	Two Parents with One Child: Dollar Gain to Parent (%)	Two Parents with One Child: Dollar Gain to Child (%)	Two Children with One Parent: Dollar Gain to Parent (%)	Two Children with One Parent: Dollar Gain to Child (%)
25,000	14.4	34.2	2.3	24.8
30,000	23.2	20.4	16.9	18.9
35,000	32.0	10.6	31.5	14.6
40,000	40.8	3.2	46.1	11.5

NOTE.—The calculations assume equal consumption by all family members. Initial wealth of parent or parents is $20,000. $R = 1.01$, $\alpha = .99$, and $\gamma = 0.75$.

full annuities; for the child the arrangement substitutes by 45.4 percent for full annuities. The last two columns of table 3 present the case of two children contracting with one parent. When each child contributes $35,000, the gain to the parent is 31.5 percent, while each 30-year-old child enjoys a 14.6 percent gain relative to consuming as a single person. The numerical differences in the table for the two different types of families reflect, on the one hand, different monetary contributions of parents relative to children and, on the other hand, differences in the rate at which resources are consumed when all family members are alive. Resources are consumed at a slower rate in the two-children–one-parent case than in the one-child–two-parent case, since each individual's survival probabilities are given equal weight in determining the optimal rate of consumption.

Is Marrying People of Similar Ages More Efficient?

Suppose one had to decide how to pair up four people, two who are old and two who are young. Is it more efficient to marry the old people together and the young people together than it is to mix ages? Intuitively, marrying two 90-year-olds together and two 20-year-olds together leaves a large chance that both 90-year-olds will die in the immediate future, and resources that they have failed to consume will be lost to the 20-year-olds who, on average, will still be alive.[10] The countervailing argument against mixed-age marriages is that mixing

[10] We assume here that any involuntary bequests that arise from the simultaneous death of both marriage partners or from the death of a surviving spouse are not inherited by any of these four individuals. Again the government can be thought of as collecting these residual bequests and distributing them each year to the newborn. We thank Finis Welch for a helpful discussion on this section.

ages involves greater risk to one of the two partners; the utility cost of this greater risk may exceed the utility gain from the increase in expected resources arising in mixed marriages.[11]

We investigated potential efficiency gains from mixed marriages between two 55-year-olds and two 30-year-olds, where each individual was risk averse at the 0.75 level. The 55-year-olds were assigned male survival probabilities, while the 30-year-olds were assigned female survival probabilities. When risk aversion equals 0.75, weights of 1.7 for the old person yield utility levels for both old and young which exceed those in the old-old, young-young marriages of table 2. The additional dollar gain to the old person from this weighted marriage with the young person is 3.1 percent; the added gain to the young person is 1.6 percent.

These additional gains from mixed-age marriages require, however, a fairly skewed distribution of consumption within the marriage. For this example the young-old weighting scheme that dominates old-old, young-young coupling involves the older spouse's consuming about 86 percent more than the younger spouse while they are both alive. If it is too costly to negotiate such an arrangement within the marriage or if the type of consumption (e.g., housing) within marriage is nonexcludable, then equal consumption marriages of individuals with similar survival probabilities (of similar ages) will be the rule rather than the exception. Of course, we have been discussing here marriages in which each spouse has the same initial dowry. The old-young marriages can dominate old-old, young-young marriages even under an equal consumption arrangement provided the dowry of the young spouse sufficiently exceeds that of the old spouse.

III. Enforcement with and without Altruism

In the absence of altruism would family members voluntarily maintain these implicit contracts as family members age? The answer is

[11] To see this consider an old-young marriage in which the young person promises to consume less than the old person in the state of nature in which both spouses survive. Suppose that this promise to the old person of higher consumption in the "both survive" state is large enough to exactly compensate the old person for the loss in expected utility from the state in which his spouse dies but he survives. The old person's expected utility from this latter "bequest" state is lower when he marries someone young, rather than someone old, because the probability of the young person actually dying is smaller. While the old person is by assumption no worse off in this compensated old-young marriage, the young person could be worse off than if he had married someone young. By entering into the compensated old-young marriage, the young person reduces his payoff in the both survive state while leaving the payoff in the bequest state unchanged. He also increases the probability of the bequest state and lowers the probability of the both survive state. Although expected consumption for the young spouse rises, the spreading of the payoffs may lower expected utility, depending on the young spouse's degree of risk aversion.

that there always are ways of structuring payments to individuals within the family so that each individual at each moment in time has a selfish interest in maintaining the original implicit contract. An equal consumption marriage contract between two individuals with the same survival probabilities and the same initial endowment is a good first example. If each spouse maintains control over his own wealth while both spouses are alive and consumes at the same rate as the other spouse, then each will separately have an incentive to continue the contract at each point in time. A similar type of individual control can be maintained in family arrangements; rather than have the parents use up all their resources before the children begin contributing to their support, the children can contribute each period in return for that period's expected parental bequest. This scenario of parents' maintaining control over their wealth until the very end, as enforcement leverage over their children, may partly explain the limited use of gifts as a tax-saving intergenerational transfer device.

The proof of this proposition is immediate from equation (10). Given their initial endowments at time $t - 1$, W^H_{t-1} and W^S_{t-1}, family members choose a value θ^* such that the contract to consume the contingent plan $[C^H_{t-1}(\theta^*), C^S_{t-1}(\theta^*), C^H_t(\theta^*), C^S_t(\theta^*), \ldots, C^H_D(\theta^*), C^S_D(\theta^*)]$ is in the core at time $t - 1$. The consumption plan at time t $[C^H_t(\theta^*), C^S_t(\theta^*), \ldots, C^H_D(\theta^*), C^S_D(\theta^*)]$ represents the period t Pareto-efficient contract corresponding to the initially chosen utility weight θ^*. This plan is in the core for a set S_t of individual endowments of family members in period t, W^H_t and W^S_t, which satisfy $W^H_t + W^S_t = W_t$. To insure that the initial contract remains a core allocation for each family member, side payments are made at time $t - 1$ when consumption in period $t - 1$ occurs. The side payments leave the period t individual endowments in set S_t. Since the initial contract $[C^H_{t-1}(\theta^*), C^S_{t-1}(\theta^*), C^H_t(\theta^*), C^S_t(\theta^*), \ldots, C^H_D(\theta^*), C^S_D(\theta^*)]$ is in the core, each selfish family member will have a personal incentive to make or accept these side payments.

There are two additional questions of enforcement to consider. One problem is that a spouse may covertly name a third party as beneficiary in his will in exchange for the same commitment by the third party or in exchange for a particular service. A second type of cheating may occur when one or both spouses covertly consume in excess of the consumption levels dictated by an optimal implicit marriage contract; while each spouse may correctly believe that he or she is the beneficiary in the other spouse's will, each may try to take advantage of the other by increasing his own consumption and thus reducing the potential bequest available to the other spouse.

These two types of cheating will be more problematic for implicit incomplete annuity agreements between friends or relatives who are

physically separated. The consumption cheating scenario can be modeled as a Nash equilibrium in which each partner chooses his consumption path by taking the other partner's consumption path and potential bequest path as given. Resources are consumed at a faster rate in the Nash equilibrium as each partner fails to consider how his consumption will diminish his expected bequest and thus the expected utility of his partner.

Using male survival probabilities we calculated for two 55-year-olds the dollar equivalent utility gain from engaging in a Nash consumption-cheating partnership. The gains in the Nash equilibrium proved to be almost identical to those in the more efficient marriage contract. For levels of risk aversion of 0.75, 1.25, and 1.75, the percentage dollar increments are, respectively, 19.9, 22.2, and 23.5. While the rate of consumption is faster in the Nash equilibrium, it is not much faster than in the marriage contract. Intuitively cheating by overconsuming is fine provided one's partner actually dies; but if one's partner survives, the early excessive consumption will require relative deprivation later on. Apparently this latter consideration dominates the former, leaving utility in the cheating equilibrium at essentially the same level as under a marriage contract. These examples suggest that consumption cheating does not represent a substantial impediment to consumption–risk-sharing arrangements.

Another means of enforcing these implicit contracts is simply altruism. All of our calculations have involved maximizing a weighted sum of individual family members' utilities. If, however, each family member is altruistic toward each other and each weights each family member's utility from consumption in the same way, then all family members would unanimously agree on the utility maximand. The calculations we have presented can, therefore, be thought of as resulting from the maximization of an agreed-upon altruistic family utility function. Since all family members agree on the maximand, there is no problem of enforcement.

IV. Summary and Conclusion

This paper has demonstrated that consumption and bequest-sharing arrangements within marriage and larger families can substitute to a large extent for complete and fair annuity markets. In the absence of such public markets, individuals have strong economic incentives to establish relationships which provide risk-mitigating opportunities. Within marriages and families there is a degree of trust, information, and love which aids in the enforcement of risk-sharing agreements. Our calculations indicate that pooling the risk of death can be an important economic incentive for family formation; the paper also

suggests that the current instability in family arrangements may, to some extent, reflect recent growth in pension and social security public annuities. The methodological approach of this paper can be applied to the study of family insurance against other types of risks. Of chief interest are those types of risks that are handled very poorly by anonymous public markets. Disability insurance and insurance against earnings losses are good examples.

Our approach has been to compare family insurance with perfect insurance. It would seem worthwhile to compare family insurance with public market insurance where the market insurance is subject to adverse selection and moral hazard problems and family insurance is not. Realistic specification of the degree of adverse selection and moral hazard may indicate that family insurance dominates public market insurance even in small families.

Finally, the paper suggests the empirical difficulty of determining whether intergenerational transfers reflect altruism or simply risk-mitigating arrangements of essentially selfish individuals in the absence of perfect insurance markets. Distinguishing between the selfish and altruistic models is fundamental to a number of major economic questions, including the impact of the social security system on national saving and the effectiveness of fiscal policy.[12]

Appendix

Computational Algorithm for the Two-Family-Members Dynamic Risk-pooling Problem

This Appendix indicates the algorithm used to solve the two-family-members dynamic programming problem, copied here as equation (A1). The algorithm for the case of three family members is similar to that for two members and is available from the authors. While we consider the iso-elastic family of utility functions, our algorithm can be applied to any homothetic utility function.

$$\begin{aligned} V_{t-1}(W_{t-1}) = \max_{W_t, C^H_{t-1}, C^S_{t-1} \geqslant 0,\, t=T,\ldots,1} & [u^H(C^H_{t-1}) + \theta u^S(C^S_{t-1}) + \alpha P_{t/t-1} Q_{t/t-1} V_t(W_t) \\ & + \alpha P_{t/t-1}(1 - Q_{t/t-1}) H_t(W_t) + \theta\alpha Q_{t/t-1}(1 - P_{t/t-1}) S_t(W_t)], \end{aligned} \tag{A1}$$

subject to

$$W_t/R + C^H_{t-1} + C^S_{t-1} = W_{t-1}. \tag{A2}$$

Again, the letters H and S correspond to the two family members with respective conditional survival probabilities $P_{t/t-1}$ and $Q_{t/t-1}$. The expression W_t is joint family wealth, θ is the weighting factor, and $H_t(W_t)$ and $S_t(W_t)$ are the expected utility levels for each family member if he alone survives to period t.

[12] See Barro 1974.

Optimal values for C_t^H and C_t^S are found recursively starting at period T and proceeding to period 0. We demonstrate that $V_t(W_t)$ may be written in the form:

$$V_t(W_t) = v_t \frac{W_t^{1-\gamma}}{1-\gamma}, \tag{A3}$$

where v_t is a constant. We also show that total family consumption, C_t, is given by

$$C_{t-1} = W_{t-1} \frac{v_T^{1/\gamma}}{v_T^{1/\gamma} + (\alpha K_t R)^{1/\gamma} R^{-1}}, \tag{A4}$$

where K_t is another constant. Given total family consumption, consumption of the two members is

$$C_{t-1}^H = \frac{C_{t-1}}{1+\theta^{1/\gamma}},\ C_{t-1}^S = C_{t-1}\frac{\theta^{1/\gamma}}{1+\theta^{1/\gamma}}. \tag{A5}$$

We demonstrate that K_t is a function of v_t and that v_{t-1} is a function of K_t. Starting then at the initial value for K_t, K_{T+1}, we can compute v_T; v_T in turn gives K_T, which in turn gives v_{T-1}. Proceeding in this fashion to period zero we compute the entire sequence of v_t's and K_t's. These values can then be used in equation (A4) to compute the ratio of consumption to wealth at each period. These ratios together with an initial level of wealth plus (A2) and (A5) generate the optimal consumption path. The homotheticity of the utility function permits us to calculate recursively the shape of the consumption path independently of the initial level of wealth.

Starting with period T the maximization problem for equation (A1) is

$$V_T(W_T) = \max \frac{1}{1-\gamma}(C_T^H)^{1-\gamma} + \theta \frac{1}{1-\gamma}(C_T^S)^{1-\gamma}$$
$$\text{s.t.}\quad C_T^H + C_T^S \leq W_T,\quad C_T^H, C_T^S \geq 0.$$

Solving this maximization and computing the indirect utility function for V_T, we have

$$V_T(W_T) = v_T \frac{1}{1-\gamma} W_T^{1-\gamma},\ \text{where } v_T = (1+\theta^{1/\gamma})^{\gamma}, \tag{A6}$$

$$C_T^H = W_T \frac{1}{1+\theta^{1/\gamma}},\ C_T^S = W_T \frac{\theta^{1/\gamma}}{1+\theta^{1/\gamma}}. \tag{A7}$$

For $t < T$, (A1) for the iso-elastic case is written as

$$\begin{aligned} V_{t-1}(W_{t-1}) &= \max_{C_{t-1}^H, C_{t-1}^S} \frac{1}{1-\gamma}(C_{t-1}^H)^{1-\gamma} + \frac{\theta}{1-\gamma}(C_{t-1}^S)^{1-\gamma} \\ &+ \alpha \frac{P_t}{P_{t-1}} \frac{Q_t}{Q_{t-1}} v_t \frac{1}{1-\gamma} W_t^{1-\gamma} + \alpha \frac{P_t}{P_{t-1}}\left(1 - \frac{Q_t}{Q_{t-1}}\right) h_t \frac{1}{1-\gamma} W_t^{1-\gamma} \\ &+ \theta\alpha\left(1 - \frac{P_t}{P_{t-1}}\right)\frac{Q_t}{Q_{t-1}} s_t \frac{1}{1-\gamma} W_t^{1-\gamma} \end{aligned} \tag{A8}$$

$$\text{s.t.}\quad C_{t-1}^H + C_{t-1}^S + \frac{W_t}{R} = W_{t-1}.$$

In going from (A1) to (A8) we use the fact that $H_t(W_t) = h_t[W_t^{1-\gamma}/(1-\gamma)]$ and $S_t(W_t) = s_t[W_t^{1-\gamma}/(1-\gamma)]$ for the iso-elastic utility function. The values for h_t and s_t are implicitly defined as the bracketed term in equation (7) in the text with $j = 0$ corresponding to time t and with each family member's survival probabilities from time t substituting for P_j.

It is easy to see from (A8) that for given total family consumption, C_t, C_t^H and C_t^S will always satisfy (A5). Hence we may rewrite (A8) as

$$V_{t-1}(W_{t-1}) = \max_{C_{t-1}} v_T \frac{1}{1-\gamma} C_{t-1}^{1-\gamma} \tag{A9}$$

$$+ \alpha \frac{W_t^{1-\gamma}}{1-\gamma}\left[\frac{P_t}{P_{t-1}} \frac{Q_t}{Q_{t-1}} v_t + \frac{P_t}{P_{t-1}}\left(1 - \frac{Q_t}{Q_{t-1}}\right)h_t + \theta\left(1 - \frac{P_t}{P_{t-1}}\right)\frac{Q_t}{Q_{t-1}} s_t\right].$$

Denoting the term in brackets by K_t we now have

$$V_{t-1}(W_{t-1}) = \max_{C_{t-1}} v_T \frac{1}{1-\gamma} C_{t-1}^{1-\gamma} + \alpha \frac{1}{1-\gamma} W_t^{1-\gamma} K_t \tag{A10}$$

$$\text{s.t.} \quad C_{t-1} + \frac{W_t}{R} = W_{t-1}.$$

Maximizing (A10) and computing the indirect utility functions yields

$$v_{t-1} = [v_T^{1/\gamma} + (\alpha R K_t)^{1/\gamma} R^{-1}]^{\gamma}, \tag{A11}$$

$$C_{t-1} = W_{t-1} \frac{v_T^{1/\gamma}}{v_T^{1/\gamma} + (\alpha K_t R)^{1/\gamma} R^{-1}}. \tag{A12}$$

References

Barro, Robert J. "Are Government Bonds Net Wealth?" *J.P.E.* 82, no. 6 (November/December 1974): 1095–1170.

Barro, Robert J., and Friedman, James W. "On Uncertain Lifetimes." *J.P.E.* 85, no. 4 (August 1977): 843–49.

Becker, Gary S.; Landes, Elisabeth M.; and Michael, Robert T. "An Economic Analysis of Marital Instability." *J.P.E.* 85, no. 6 (December 1977): 1141–87.

Friend, Irwin, and Blume, Marshall E. "The Demand for Risky Assets." *A.E.R.* 65 (December 1975): 900–922.

Kotlikoff, Laurence J., and Spivak, Avia. "The Family as an Incomplete Annuities Market." Working Paper no. 362, Nat. Bur. Econ. Res., June 1979.

Schultz, Theodore W., ed. *Economics of the Family: Marriage, Children, and Human Capital.* Chicago: Univ. Chicago Press (for Nat. Bur. Econ. Res.), 1974.

Sheshinski, Eytan, and Weiss, Yoram. "Uncertainty and Optimal Social Security." *Q.J.E.* (1981), in press.

U.S. Department of Health, Education, and Welfare, Social Security Administration. *United States Population Projection for OASDHI Cost Estimates.* Actuarial Study no. 62. Washington: U.S. Department of HEW, Social Security Administration, December 1966.

Yaari, Menahem E. "Uncertain Lifetime, Life Insurance, and the Theory of the Consumer." *Rev. Econ. Studies* 32 (April 1965): 137–50.

An Economic Theory of Self-Control

Richard H. Thaler

Cornell University

H. M. Shefrin

University of Santa Clara

The concept of self-control is incorporated in a theory of individual intertemporal choice by modeling the individual as an organization. The individual at a point in time is assumed to be both a farsighted *planner* and a myopic *doer*. The resulting conflict is seen to be fundamentally similar to the agency conflict between the owners and managers of a firm. Both individuals and firms use the same techniques to mitigate the problems which the conflicts create. This paper stresses the implications of this agency model and discusses as applications the effect of pensions on saving, saving and the timing of income flows, and individual discount rates.

For many years Christmas clubs paid no interest. Members deposited money each week but could only withdraw the money on December 1. The clubs were very popular, although they seemed to be dominated by simply depositing money in an interest-bearing savings account.

Passbook loans do still exist. These loans allow an individual with $5,000 in a savings account earning 5 percent interest to borrow at 9 percent using the balance as collateral, instead of at 10 percent with

The authorship of this paper was fully collaborative. The order of the names on this paper and its companion (Shefrin and Thaler 1980) corresponds to our division of labor. This paper was written while Thaler was a visiting scholar at the National Bureau of Economic Research at Stanford, California. While there he received financial support from the Kaiser Family Foundation and the Robert Wood Johnson Foundation. Helpful comments on earlier drafts were provided by George Ainslee, James Buchanan, Tom Russell, and numerous other friends and colleagues. We thank them all.

[*Journal of Political Economy*, 1981, vol. 89, no. 2].

no collateral. Obviously the individual could simply withdraw the money from his savings account at an (opportunity) cost of only 5 percent.

Smoking clinics are a new and thriving business. A smoking clinic will help people who want to stop smoking—for a fee of several hundred dollars.

What does economic theory have to say about these institutions? George J. Stigler provides the following analysis:

> One can of course explain the participation in a Christmas fund by introducing another item of preference: a desire of people to protect themselves against a future lack of willpower. . . . If we stopped the analysis with this explanation, we would turn utility into a tautology: a reason, we would be saying, can always be found for whatever we observe a man to do. In order to preserve the predictive power of the utility theory, we must continue our Christmas fund analysis as follows. The foregone cost of putting money in a Christmas fund is the interest one could earn by putting the same money in a savings account. If interest rates on savings accounts rise, the cost of buying protection against a loss of willpower rises and less of it ought to be bought. . . . [Stigler 1966, p. 57]

We agree with the remarks above and therefore offer a model that says more about these institutions than the fact that the demand for their services will be negatively related to price. We do so by proposing a simple extension of orthodox models, using orthodox tools, that permits such behavior to be viewed as rational. This rationalization is based on an analysis of the technology of self-control using the theory of agency rather than reliance on ad hoc explanations in which transaction costs, taxes, and income effects play a major role. Our model can predict, in a nontautological way, the circumstances in which these kinds of behavior will be observed. In particular, our new theory of intertemporal choice has important implications for theories of saving behavior. In the last section of the paper we discuss some of these implications and offer empirical evidence in support of our ideas.

I. The Model

> The idea of self-control is paradoxical unless it is assumed that the psyche contains more than one energy system, and that these energy systems have some degree

of independence from each other. [DONALD MCINTOSH 1969]

Why individuals would impose constraints on their future behavior is a problem that has received attention from economists since Strotz's (1955–56) classic paper.[1] Strotz and those who have followed him (Pollak 1968; Blackorby et al. 1973; Peleg and Yaari 1973; Hammond 1976; and Yaari 1977) have analyzed the phenomenon as one of *changing tastes.* In Strotz's formulation a conflict occurs between today's preferences and tomorrow's preferences if the discount function used today is not exponential with a constant exponent.

Our framework differs from the changing-tastes literature in that we model man as having two sets of preferences that are in conflict at a single point in time. This idea is certainly not new. Adam Smith used a two-self model much like ours in his *Theory of Moral Sentiments* (1759). More recently Schelling (1960, 1978) and Buchanan (1975) have recognized the importance of simultaneous conflict in understanding self-control problems. Outside economics the idea is commonplace, with the writings of Freud (1958) and Berlin (1969) deserving special mention.

Nonetheless, to the best of our knowledge our work is the first systematic, formal treatment of a two-self economic man. We have adopted a two-self model because, as McIntosh says above, the notion of self-control is paradoxical without it. Furthermore, we utilize an organizational analogy that leads to both insights into human behavior and a rich explanatory model. We will briefly describe the model here. We do so in order to make explicit the nature of our two-self conceptualization. The model also leads to specific predictions about behavior, which are discussed in Section III below. Further details are available in our more formal companion paper (Shefrin and Thaler 1980).

Our model is cast in discrete time. Consider an individual with a fixed income stream $y = (y_1, y_2, \ldots, y_T)$. Think of period T as retirement and let $y_T = 0$. Let the individual choose a nonnegative level of consumption c_t in t. Call $c = (c_1, c_2, \ldots, c_T)$ a consumption plan. The conflict between short-run and long-run preferences is introduced by viewing the individual as an organization. At any point in time the organization consists of a planner and a doer.[2] The planner is concerned with lifetime utility, while the doer exists only for one period and is completely selfish, or myopic. The period t doer is assumed to have direct control over period t consumption rate c_t. The

[1] Two important contributions by noneconomists are Ainslee (1975) and Elster (1977).

[2] These terms originated in an early draft of Thaler's (1980).

doer's utility function is given by $Z_t(\cdot)$; Z_t is taken to be independent of all components of c except c_t.[3] Furthermore, suppose initially that Z_t is strictly increasing and concave in c_t.

In our model the planner does not actually consume but, rather, derives utility from the consumption of the doers. Therefore the planner's utility function is given by $V(Z_1, Z_2, \ldots, Z_T)$. Observe that a plan which maximizes V subject to the present value budget constraint, $\sum_{t=1}^{T} c_t \leq \sum_{t=1}^{T} y_t = Y$, is considered optimal from the planner's point of view. However, without some method to control the doer's actions, this plan cannot be implemented. Indeed, under the assumptions above, the actual consumption stream chosen would have total lifetime income consumed during the first period, when the period-one doer would borrow $Y - y_1$ on the "perfect" capital market. In order to prevent this from occurring, the planner requires some psychic technology capable of affecting the doer's behavior. Two main techniques are available for this: (1) The doer can be given *discretion* in which case either his *preferences* must be modified or his *incentives* must be altered, or (2) the doer's set of choices may instead be limited by imposing *rules* that change the constraints the doer faces.

We begin by analyzing the case in which no rules are used. We refer to this case as *pure discretion.* While we do not believe that the pure discretion case is empirically important (most people appear to use at least some kinds of rules), it provides a useful foundation for our model, to which rules are easily added. Furthermore, since this case corresponds closely to that usually considered in economics, it highlights the differences between our model and the standard framework.

Recall that $Z_t(\cdot)$ was assumed to be unbounded. We now specify that Z_t depends on a preference modification parameter θ_t selected by the planner. The choice of θ_t allows the planner to alter Z_t such that it possesses an internal maximum. By appropriately selecting θ_t, any desired c_t may be obtained; however, the lower the desired c_t is, the more modification will be required. Furthermore, $\partial Z_t/\partial\theta_t$ is negative; that is, modification reduces short-run utility and is therefore costly. Finally, we assume that the marginal cost of modification increases with θ. Thus successive reductions in c_t require increasing reductions in Z_t.

If the planner could exercise direct control over the choice of a consumption plan, θ_t would be set equal to zero for all t (because modification is costly) and c would be selected to maximize V subject to the budget constraint. Since under pure discretion this is assumed to

[3] The extreme assumption that doers do not care at all about past or future doers is adopted just for expositional simplicity. Other arguments of Z_t could easily be added.

be infeasible, the planner must instead choose $\theta = (\theta_1, \theta_2, \ldots, \theta_T)$ to maximize V. (The solution to the problem is described in Shefrin and Thaler [1980].) Essentially, modification is increased until the marginal utility derived from additional consumption in retirement (period T) equals the marginal loss in utility in earlier periods due to modification.

In the more general case when both rules and preference modification are permitted, the planner may also alter the budget constraint facing each doer. This allows the planner to reduce c_t without incurring modification costs. However, since available rules are imperfect (see next section), the planner will have to trade off modification costs with the opportunity costs associated with using second-best-type rules. Rules are formally incorporated into the planner-doer model in Shefrin and Thaler (1980).

II. Techniques to Reduce Conflicts in Individuals and Organizations

We have characterized self-control as an internal conflict resembling the principal-agent conflict between the owner and manager of a firm (see Ross 1973; Jensen and Meckling 1976). In this section we describe the actual techniques used by individuals and firms to mitigate these conflicts. The techniques fall into the two categories highlighted in the model: rules and incentives. We provide many illustrations of the methods individuals use because these are in essence part of our model. When individuals use rules it is impossible to characterize their behavior simply with first-order conditions. The limits on the kinds of rules which individuals will find feasible lead to the specific predictions about saving behavior discussed in Section III. In this section we also compare the techniques individuals use with those used in firms. The close correspondence we find lends intuitive support to our principal-agent model.

A. *Methods to Alter Incentives*

Individuals use three basic techniques to alter the doer's incentives. First, the doer's preferences can be modified directly. Some individuals consider saving a good in and of itself.[4] In this case doer myopia does not inhibit saving. Second, inputs to a saving or dieting program can be explicitly monitored via weekly budgets or calorie counting

[4] This idea has been suggested by Scitovsky (1976). The importance of norms in controlling individual behavior is also stressed heavily by Adam Smith (1759, p. 326) and by Irving Fisher (1930).

(customers of diet clinics and credit counselors are advised to do this). Simply keeping track seems to act as a tax on any behavior which the planner views as deviant. Third, incentives can be explicitly altered: Alcoholics take the drug Antabuse which makes them ill if they take a drink; academics agree to give a paper at a conference to provide a proximate incentive to write it.

Firms use the same three methods. First, profit-sharing plans are quite popular even though they offer only trivial financial incentives to all but the highest executives. We believe firms adopt them because they help create an atmosphere in which the employees' preferences are more similar to those of the owner. Second, firms monitor departmental inputs through cost accounting and then tie compensation to these input measures. Third, departmental profits are measured and used as performance measures for managers.

B. *Methods to Alter Opportunities: Rules*

If the costs of monitoring and persuasion are high, individuals will resort to rules that restrict the doer's opportunities. In the extreme, all doer discretion can be eliminated using what Strotz referred to as the strategy of precommitment. Such behavior is rational in our model if the rule can approximate the choices that the planner would select. Market precommitment institutions are observed, as we would predict, in such areas as saving and dieting. For example, people pay to go to "fat farms" which essentially are resorts that promise not to feed their customers.

Less extreme rules can limit the *range* of doer discretion, usually through the use of self-imposed rules of thumb. In the savings context several such rules appear to be commonly used. These rules alter the budget constraint faced by the doer in much the same way as credit limits imposed by lenders do.[5] A simple first departure from pure discretion is a ban on borrowing, the so-called debt ethic. A somewhat weaker rule which seems common is to prohibit borrowing except for specific purchases, like houses and automobiles. Another rule of thumb is a prohibition on dissaving combined with limits on borrowing. Using this rule of thumb, a person might borrow and lend simultaneously in spite of a substantial difference in the interest rates, as in the case of the passbook loan. The loan allows him to transfer consumption across time periods while it provides a regimented repayment scheme.

[5] While it is difficult to document the extent to which these precise rules are used, 85 percent of Cagan's (1965) sample of Consumers Union members reported using one of these or similar rules to determine monthly saving.

Rules can also eliminate discretion over a specific *class* of decisions for which the conflict is particularly acute. Dieters try not to keep cheesecake in the refrigerator and will refuse invitations to lavish dinner parties; problem gamblers avoid Las Vegas. Also, many smokers pay more for their cigarettes by buying them by the pack instead of the carton—it helps enforce a self-imposed ration such as one pack a day.

Again, the same types of rules are observed in organizations.[6] "Pure" rules are observed most often in bureaucratic organizations because the costs of monitoring output are so high. Rules that limit managerial discretion over a particular range are frequently in the form of guidelines (e.g., a plant manager can adapt any investment that exceeds some stated rate of return). Similarly, rules may prohibit discretion over a specific class of decisions; loan officers, for example, might need approval for loans to relatives or friends.

We wish to make three other points about internal rules of thumb. First, it is useful to consider these rules as learned as much as chosen. Rules like the debt ethic are learned from parents and other models, which suggests that there will be differences in the use of rules depending on social class, education, and age. Second, rules of thumb are likely to become habits. By establishing a routine, the doer decision process can be avoided. Third, to the extent that the rules do become habits, there will be rigidities built into the individual's behavior.[7] The implications of these observations are discussed in the next section.

III. Implications

We now turn to a discussion of the implications of the planner-doer model. What predictions about behavior can be made with our model that are inconsistent with the standard model? Some of these predictions are obvious. We predict that people will rationally choose to impose constraints on their own behavior. Furthermore, we predict that such precommitments will occur primarily for those goods whose benefits and costs occur at different dates. We present here some less obvious predictions based on our model.

Because our framework is richer than the standard theory, vari-

[6] For a discussion of firms imposing constraints on their future financial policies, see Myers (1977).

[7] Habits can be formally introduced in two stages. At first stage Z_t can also be made a function of θ_s, $s < t$, to reflect the fact that self-control at the early dates renders self-control at later dates less costly. At the second stage, Z_t can also be parameterized on the adopted rule. In this case the function Z_t could exhibit an internal maximum in c_t when $\theta_t = 0$. It is interesting to note that modification would now be required in order to break the rule. This would tend to explain miserliness, for instance.

ables that the standard theory treats as irrelevant differences in form we model as differences in substance. In Sections IIIA and IIIB we consider how differences in the form of payment (holding the level constant) affect saving decisions, and in Section IIIC we present variables other than borrowing and lending rates that determine individual marginal rates of time preference.

A. *Pensions and Saving*

Consider two identical individuals with the same total income and wealth. Assume that they both save some fraction s of their income. Now give one a mandatory pension plan that forces him to save $p < s$. What will happen to total saving? Though it is difficult to get a specific prediction from the standard model (see Feldstein 1977), a first-order prediction would be that total saving is unaffected. Other forms of saving should fall by the approximate amount of the pension.

Our model has a different prediction: The pension plan produces saving at no psychic cost. Modification costs occur only when saving is voluntarily withheld. This is what we call "discretionary saving." Thus since the marginal cost of saving is lower at the old saving level, we expect total saving to go up. (In other words, the offset in other saving will be less than the size of the pension.)

Furthermore, in two cases the offset will be essentially zero: (1) The individual uses a saving rule such as "save s percent of disposable income" which is not changed when the pension is introduced (total saving increases by $[1 - s]py_t$), and (2) the individual uses discretion but treats saving as a good for its own sake rather than as a transfer to future consumption. In this case discretionary saving and retirement saving are not perfect substitutes as in the standard model; in fact, their cross-elasticities of demand could be zero.[8]

The effect of pensions on saving has been investigated with individual data by Cagan (1965), Katona (1965), and Munnell (1974, 1976). All obtained similar results. Cagan used a sample of Consumers Union members. He found that for those members with a mandatory pension plan, other saving actually was higher than for those without pensions (see Cagan 1965, p. 21). Munnell (1974) replicated Cagan's study using the same data source. She used a different measure of saving, replaced before-tax income with after-tax income, and restricted her analysis to a subset of observations she thought to be more reliable. She then regressed the nonpension saving to income ratio on several variables including a pension dummy. Her basic result

[8] A formal treatment of the effect of pensions on saving appears in Shefrin and Thaler (1980).

was that the pension had no effect on other saving (i.e., a zero offset). The coefficient of the pension dummy was never significant. Its highest t value occurred for the 55–65 age group ($t = 1.2$), for which those with pensions saved 3 percent less than those without pensions. Though other explanations (such as selectivity bias) have been offered for these results, we find that they lend support to our model.

B. Saving and the Timing of Income Flows

The importance of current disposable income (as opposed to permanent income) yields another prediction from our model that differs from the standard theory. Consider two identical individuals, S and B; S receives a salary of $12,000 per year paid in 12 monthly installments of $1,000, while B receives a salary of $10,000 per year paid in monthly installments plus a guaranteed bonus of $2,000 paid in March each year. Standard theories of saving behavior would predict that these two individuals would save the same amount. Our model predicts that, on average, B will save more.

Although we know of no test of this hypothesis,[9] there is one bit of circumstantial evidence. In Japan, where there is a very high saving rate, bonus schemes are quite common. We think this is no coincidence. A test would be possible, given the right data. We predict that individuals who are paid a portion of their salary via a lump-sum bonus will have higher saving rates than those who receive their compensation in a smooth pattern. How does this follow from the model?

We have characterized saving behavior primarily as a set of self-imposed rules of thumb and externally enforced saving plans. For an individual like B, those rules and plans will be based on his regular monthly income. Contributions to pension plans, payments on whole life insurance policies, mortgage payments, and so forth must be made on a regular basis. Furthermore, most individuals prefer to have their monthly inflows and outflows roughly balance.[10] For B to act like S, he could deposit the bonus in the bank and draw it down gradually during the year (as if his salary were $12,000), or he could borrow the $2,000 over the course of the year and repay the loan when the bonus is paid. However, we feel that neither of these behavior patterns is likely to be widely observed. Notice that they violate either the ban on borrowing or the ban on dissaving. To the extent that these rules of thumb are used, imitating the behavior of S will be difficult. We

[9] However, a related issue is investigated by Landsberger (1966).

[10] This explains why many teachers sacrifice interest by electing to receive their academic-year salary paid in 12 monthly installments (September–August) rather than 10 (September–June).

believe that for the typical individual much of the bonus will end up being saved, especially through the purchase of durable goods. This amounts to the use of an auxiliary rule regarding the disposition of bonuses received. We expect that total saving for B will exceed that for S because of the technology of self-control. By paying the bonus the firm is acting as an external self-control device (much like parents who tell children that money gifts at Christmas must go into a savings account). Temptations to spend during the year will be overcome because the smaller monthly salary will make them seem beyond the individual's means. This technology seems to have been recognized by the millions of taxpayers each year who claim too few exemptions in order to assure a tax refund.[11] Obviously some self-deception (doer deception) is necessary for this device to work, but doers can apparently be deceived quite easily. How else can one explain the not uncommon practice of knowingly setting one's watch a few minutes ahead ("in order to get to places on time")?

As the analysis in this section suggests, the shape of the income stream will affect the type of saving strategy adopted. Those individuals with variable and uncertain incomes will find a discretionary rule such as saving s percent each month difficult to enforce in the low-income months. Without a mandatory saving plan, they would have to adopt some more complex strategy to save effectively. Similarly, those individuals whose incomes are expected to decline (such as professional athletes) would prefer to save a large proportion of their high current incomes and a smaller (perhaps negative) proportion of their lower future incomes. For both, a mandatory pension plan is particularly likely to increase total saving. A more sophisticated analysis of the effect of pensions on saving might detect differences of this sort.

The case of athletes points up the extreme differences in behavior that self-control can produce. Their declining income stream creates a difficult self-control problem in the high-income years. Some athletes hire agents to invest their incomes and limit their current spending, and many of them become rich. Others rely on discretionary strategies and end up bankrupt. Both types of behavior are possible in our model, depending on the degree of planner control and the types of precommitment strategies available. The best predictors of which individuals will fall into which groups are probably related to family background, since the family is the most likely place for the individual to learn (or not learn) the rules and norms necessary to overcome the self-control problems.[12]

[11] E.g., in 1969, 55 million taxpayers received refunds while 18 million owed taxes. In dollars, overpayments exceeded underpayments by 39 percent.

[12] The last two paragraphs were prompted by a suggestion from Sam Peltzman. Irving Fisher also discussed some of these issues. He notes that some individuals spend

C. *Individual Marginal Rates of Time Preference*

The orthodox theory of intertemporal choice as formulated by Irving Fisher produces a very strong result. Each person should equate his marginal rate of time preference (MRTP) with the relevant after-tax interest rate.[13] Thus the theory predicts that all individuals who face the same after-tax interest rate will make the same marginal intertemporal choices: Specifically, they will act as if they used the after-tax interest rate as their discount rate. This follows because individuals are assumed to use capital markets to arbitrage away any difference between what Fisher called their "rate of impatience" and the interest rate. It should be noted, however, that this result can fail to hold if capital markets impose quantity constraints on borrowers (capital rationing). In this case borrowers may be forced to stop borrowing even though their rate of impatience exceeds the interest rate.

In exactly the same fashion, self-imposed borrowing constraints such as those discussed in Section IIB prevent the complete internal arbitrage from taking place. Thus in our model, in which such constraints play an important role, the presumption that individual MRTPs will equal the interest rate no longer holds. Indeed, we expect to observe behavior that implies an MRTP greater than the interest rate and at the same time an unwillingness to engage in additional borrowing. Two points need to be raised about this implication. First, once rules are incorporated into the analysis, the failure to equate the MRTP to the interest rate may not violate an optimality condition. If rules are used it is because they lead to higher levels of utility than pure discretion would. The inequality created is costly, but the cost arises from the necessity of using a second-best technology. Because rules by nature must be simple, rules that select the precisely correct consumption bundle in every situation are infeasible. Second, notice that if a quantity constraint is binding, whether internally or externally imposed, observed MRTPs will be equal to or greater than the interest rate.

Attempts to measure individual MRTPs appear in studies by Kurz, Spiegelman, and West (1973) and Hausman (1979). Hausman studied families' purchases of room air conditioners. The trade-off between initial outlay and operating costs permitted him to estimate implicit MRTPs. The mean MRTP in his sample was about 25 percent, clearly above any relevant interest rate. Kurz et al. obtained similar results

---

their weekly paycheck at the "grog house." (Others, we believe, avoid the grog house precisely on those days.) He attributes much of the observed differences in behavior to social class. Our differences with Fisher are discussed in the next section.

[13] We define the MRTP to be the marginal rate of substitution between tomorrow's consumption and today's consumption minus one.

by asking hypothetical questions of participants in the Seattle and Denver Income Maintenance Experiments. They asked a sample of participants a series of questions of the following sort: What size bonus would you demand today rather than collect a bonus of $100 in 1 year? Several different forms of this type of question were asked, and the results were striking. For whites the mean rate of time preference implied by their answers varied between 36 and 76 percent. For blacks the rates varied between 40 and 122 percent. Of particular interest for present purposes is the fact that this sample included only those respondents who said they could borrow either $500 to make an installment purchase or $1,000 in cash. Furthermore, 81.3 percent of this subsample reported that they would not borrow $1,000 at current interest rates. (The mean perceived rate was generally less than 20 percent.) This strongly suggests the use of self-imposed borrowing constraints.

Once individual differences in MRTPs are anticipated, it becomes interesting to ask what factors determine those differences. A detailed examination is beyond the scope of this paper, but the model predicts that those factors determining individual rates of impatience will (for the reasons stated above) also affect observed MRTPs. These factors are discussed at length by Fisher (who draws on the writings of John Rae and Eugen Bôhm-Bawerk). Fisher believed that age, income, and marital status affect the rate of impatience (in obvious directions). He felt that the shape of the income stream as well as the level was important; if income were expected to rise, the rate of impatience would be higher. Six "personal characteristics" were also deemed to be important: foresight, self-control, habit, expectation of life, concern for the lives of other persons, and fashion.[14]

In principle, the planner-doer model can be tested against the standard model quite simply. Observe some intertemporal choice that implies a discount rate (such as Hausman's study of air conditioners). Then regress the implied discount rate on the factors above *and* the individual's borrowing rate. The standard framework implies that only the borrowing rate will be a significant predictor. Our model implies that the other factors will also be important. Hausman did test one such variable. The implied rate of discount was computed for six income classes. His results appear in table 1. Clearly the variation in discount rates cannot be attributed solely to variations in borrowing rates.

We would also expect age and social class to be important in predicting individual intertemporal choices. The young behave impatiently in part because they have yet to master the techniques of

[14] Fisher 1930, p. 81. On this topic also see Maital and Maital (1977).

TABLE 1

Estimated Discount Rates Using Mean Population Estimates

Income Class ($/yr)	Observations (N)	Implied Discount Rate (%)
600	6	89
10,000	15	39
15,000	16	27
25,000	17	17
35,000	8	8.9
50,000	3	5.1

Source.—Hausman 1979.

self-control. To the extent that these techniques are learned from parents, class differences will be observed. Most important, our model stresses the theoretical admissability of these variables. Only further empirical work can establish their relative explanatory power.

IV. Summary and Conclusion

We now briefly recapitulate our argument. We have investigated intertemporal choice as a problem in the economic theory of self-control. As the quotation at the beginning of Section I states, the concept of self-control is paradoxical unless some kind of multiself model of man is adopted. We have introduced self-control into a formal model of intertemporal choice by modeling man as an organization with a planner and many doers. Conflict occurs because the doers are myopic (i.e., selfish). This conflict is fundamentally similar to the agency relationship between the employer and the employee, and individuals use many of the same strategies that organizations adopt to deal with their "conflicts of interest." These strategies can involve doer/employee discretion while their incentives have somehow been altered, or they may entail the implementation of precommitment (a rule) to avoid the doer/employee decision process altogether.

The close correspondence between the solutions to control problems adopted by organizations and individuals provides strong support for our model. Although our model is nontraditional, our tools are strictly traditional. Formally, our model closely resembles that used by Ross (1973) in his study of the theory of agency. Finally, we note that ours is a theory of rational behavior, just as Ross's theory is of profit-maximizing behavior.

Many applications of the model are possible, and we have discussed a few briefly. The most important applications are in the study of

individual saving behavior. Our hypotheses are quite different in spirit from the permanent income and life-cycle hypotheses that currently dominate the literature. On the basis of the evidence presented here, we feel these theories of saving should be reevaluated.

References

Ainslee, George. "Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control." *Psychological Bull.* 82 (July 1975): 463–96.

Berlin, Isaiah. "Two Concepts of Liberty." *Four Essays on Liberty.* New York: Oxford Univ. Press, 1969.

Blackorby, Charles; Nissen, David; Primont, Daniel; and Russell, R. Robert. "Consistent Intertemporal Decision Making." *Rev. Econ. Studies* 40 (April 1973): 239–48.

Buchanan, James. "The Samaritan's Dilemma." In *Altruism, Morality, and Economic Theory,* edited by Edmund S. Phelps. New York: Russell Sage, 1975.

Cagan, Philip. *The Effect of Pension Plans on Aggregate Saving: Evidence from a Sample Survey.* New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1965.

Elster, Jon. "Ulysses and the Sirens: A Theory of Imperfect Rationality." *Soc. Sci. Information* 16 (October 1977): 469–526.

Feldstein, Martin. "Do Private Pensions Increase National Saving?" Working Paper no. 186, Nat. Bur. Econ. Res., Cambridge, Mass., 1977.

Fisher, Irving. *The Theory of Interest.* London: Macmillan, 1930.

Freud, Sigmund. "Beyond the Pleasure Principle." In *The Standard Edition of the Complete Psychological Works of Sigmund Freud,* edited by James Strachey and Anna Freud. London: Hogarth, 1958.

Hammond, Peter J. "Changing Tastes and Coherent Dynamic Choice." *Rev. Econ. Studies* 43 (February 1976): 159–73.

Hausman, Jerry A. "Individual Discount Rates and the Purchase and Utilization of Energy-using Durables." *Bell J. Econ.* 10 (Spring 1979): 33–54.

Jensen, Michael C., and Meckling, William H. "Theory of the Firm: Managerial Behavior, Agency Costs, and Ownership Structure." *J. Financial Econ.* 3 (October 1976): 305–60.

Katona, George. *Private Pensions and Individual Saving.* Ann Arbor: Univ. Michigan Inst. Soc. Res., 1965.

Kurz, M.; Spiegelman, R.; and West, R. "The Experimental Horizon and the Rate of Time Preference for the Seattle and Denver Income Maintenance Experiments: A Preliminary Study." Menlo Park, Calif.: SRI Internat. Res. Memorandum no. 21, November, 1973.

Landsberger, Michael. "Windfall Income and Consumption: Comment." *A.E.R.* 56 (June 1966): 534–40.

McIntosh, Donald. *The Foundations of Human Society.* Chicago: Univ. Chicago Press, 1969.

Maital, Schlomo, and Maital, Sharona. "Time Preference, Delay of Gratification and the Intergenerational Transmission of Economic Inequality: A Behavioral Theory of Income Distribution." In *Essays in Labor Market Analysis: In Memory of Yochanan Peter Comay,* edited by Orley C. Ashenfelter and Wallace E. Oates. New York: Wiley, 1977.

Munnell, Alicia H. *The Effect of Social Security on Personal Saving.* Cambridge, Mass.: Ballinger, 1974.

———. "Private Pensions and Saving: New Evidence." *J.P.E.* 84, no. 5 (October 1976): 1013–32.

Myers, Stewart C. "Determinants of Corporate Borrowing." *J. Financial Econ.* 5 (November 1977): 147–75.

Peleg, Bezalel, and Yaari, Menahem E. "On the Existence of a Consistent Course of Action When Tastes Are Changing." *Rev. Econ. Studies* 40 (July 1973): 391–401.

Pollak, Robert A. "Consistent Planning." *Rev. Econ. Studies* 35 (April 1968): 201–8.

Ross, Stephen A. "The Economic Theory of Agency: The Principal's Problem." *A.E.R. Papers and Proc.* 63 (May 1973): 134–39.

Schelling, Thomas C. *The Strategy of Conflict.* Cambridge, Mass.: Harvard Univ. Press, 1960.

———. "Egonomics, or the Art of Self-Management." *A.E.R. Papers and Proc.* 68 (May 1978): 290–94.

Scitovsky, Tibor. *The Joyless Economy: An Inquiry into Human Satisfaction and Consumer Dissatisfaction.* New York: Oxford Univ. Press, 1976.

Shefrin, H. M., and Thaler, Richard H. "Rules and Discretion in a Two-Self Model of Intertemporal Choice." Mimeographed. Ithaca, N.Y.: Cornell Univ., April 1980.

Smith, Adam. *Theory of Moral Sentiments.* London: Millar, 1759.

Stigler, George J. *The Theory of Price.* 3d ed. New York: Macmillan, 1966.

Strotz, Robert H. "Myopia and Inconsistency in Dynamic Utility Maximization." *Rev. Econ. Studies* 23, no. 3 (1955–56): 165–80.

Thaler, Richard H. "Toward a Positive Theory of Consumer Choice." *J. Econ. Behavior and Org.* 1 (1980): pp. 39–60.

Yaari, Menahem E. "How to Eat an Appetite-arousing Cake." Research Memorandum no. 26, Center Res. Math. Econ. and Game Theory, Hebrew Univ., June 1977.

Comments

A Note on Loss of Control and the Optimum Size of the Firm

Antonio Camacho and William D. White

University of Illinois at Chicago Circle

This note has been inspired by a recent paper in this journal by Calvo and Wellisz (1978). Its main purpose is to suggest that, in constructing models to study the size of firms, more attention needs to be given to the conditions which bring about their existence.

In Section II of their article, Calvo and Wellisz prove a proposition that given certain assumptions about supervision, if a firm of any size exists, then loss of control will not impose finite limits on its size. Although their result is correct, they, as well as some other researchers working in this field, fail to give explicit reasons for firms to exist. However, as we will argue below, there are not convincing reasons, under their assumptions, for firms of any size to exist, thus making their case for the appearance of unbounded firms rather weak.

This result prompts us to suggest, as stated above, that more attention must be given in this area of research to the very conditions which bring about the existence of firms. For instance, models which, in the spirit of Williamson's (1967) paper, explicitly consider problems with information and communication, or models where it is shown that the activities of workers or agents may be coordinated more efficiently through extra market arrangements, may both produce realistic and interesting results regarding the size of firms.

We proceed as follows. First, we prove that, under Calvo and Wellisz's assumptions, firms will never exist if workers can be self-employed. Second, we show by means of a "setup" cost example that,

We are grateful to Joseph Persky for his comments on an earlier version of this note. Any errors which remain are, of course, our own.

[*Journal of Political Economy*, 1981, vol. 89, no. 2]

even if conditions which limit the opportunities of workers for self-employment exist, Calvo and Wellisz's assumptions regarding skills, utility functions, and productivity of workers seem to imply that there are more efficient arrangements to carry out production activities than the formation of firms. The note concludes with an intriguing conjecture and a suggestion.

I. No Limitations on Opportunities for Self-Employment

We will proceed now to prove that, if there are no limitations on workers' opportunities for self-employment, then under Calvo and Wellisz's assumptions firms will never exist. We will obtain this result by showing that, under their assumptions, a worker can always achieve at least as high a level of utility by working for himself as he can by working as an employee for any firm which makes a positive profit out of hiring him.

In their model, Calvo and Wellisz assume that the (Von Neumann-Morgenstern) utility index, U, of every worker is the same and depends on consumption, c, and effort, e; more specifically they assume:

$$U = u(c) - v(e), c \geqslant 0,\ 0 \leqslant e \leqslant 1, \tag{1}$$

where $u' \geqslant 0, v' \geqslant 0, u'' \leqslant 0$, and $v'' \leqslant 0$. (The last two conditions rule out risk-loving behavior.) Effort is assumed to be at a maximum when e equals one, while the worker is assumed to be completely idle when e is equal to zero.

Workers employed in firms are subject to monitoring where P is the probability that a worker will be detected shirking. Workers who are not detected shirking are paid a wage of w. Workers who are found shirking are paid we.

For any P and w, a worker will select that level of effort which maximizes his expected utility z:

$$z = P[u(we) - v(e)] + (1 - P)[u(w) - v(e)]. \tag{2}$$

Let $\bar{e}$ be the value of e which maximizes z.

Calvo and Wellisz assume also that the productivity of every worker is the same and equal to βe $(0 \leqslant e \leqslant 1)$ when working at the level of effort e. Given this assumption, a firm's expected revenue from employing a worker at a wage w and with a level of supervision P will be $\beta\bar{e}$. The expected cost to the firm of employing the worker will be the expected payments to the worker, $P(w\bar{e}) + (1 - P)w$, plus the expected cost of providing a level of supervision P. Thus, the total expected cost of employing a worker will be greater than or equal to $P(w\bar{e}) + (1 - P)w$. A necessary condition for the firm to make a positive expected profit is then

$$\beta\bar{e} > P(w\bar{e}) + (1 - P)w$$

or

$$w < \frac{\beta\bar{e}}{P(\bar{e} - 1) + 1}. \tag{3}$$

Let $E(\hat{U})$ be the maximum expected utility which a worker can obtain by working for a firm at a wage rate w and under a level of supervision P:

$$E(\hat{U}) = P[u(w\bar{e}) - v(\bar{e})] + (1 - P)[u(w) - v(\bar{e})]. \tag{4}$$

The expected utility of a worker when self-employed and applying himself to his work at the same level of effort $\bar{e}$ as in (4) is

$$E(U) = u(\beta\bar{e}) - v(\bar{e}). \tag{5}$$

Taking into consideration condition (1) and inequality (3), we can easily obtain

$$E(\hat{U}) \leqslant u(\beta\bar{e}) - v(\bar{e}),$$

which in view of (5) proves our assertion.

II. Limitations on Opportunities for Self-Employment

We will now discuss conditions that may limit workers' opportunities to be self-employed. To facilitate this discussion we will first enumerate the main assumptions of Calvo and Wellisz's model: Everyone has the same skills; everyone has the same utility function and, therefore, the same attitude toward risk; there is no specialization of labor, no need for communication of instructions or for coordination of activities; there are constant returns to scale when everyone applies the same level of effort to production activities.

Given these assumptions and the additional assumption that there are no institutional restrictions on self-employment, it appears that the only condition which might limit opportunities for self-employment for part of the population is the existence of some factor of production, say land, whose ownership is unevenly distributed. This situation might allow owners to pay workers (those who lack any endowment of this factor) a wage below their marginal product. But let us analyze this situation in more detail. Consider two cases: (i) only one individual owns all the land; (ii) more than one individual owns land.

In case i, an individual who is a monopolist may be able to start a firm and make total profits greater than β by hiring workers at a wage below their marginal product. Then, as Calvo and Wellisz rightly prove, the owner can increase his profits by adding layers to the hierarchical structure of his firm until he employs everybody.

But this is a rather cumbersome arrangement. Suppose that the total amount of land is L, the total number of workers is k, and the maximum profit the monopolist can extract by employing k workers is $\pi(k)$. By renting each worker a plot of land L/k at a price of $\pi(k)/k$, the monopolist can obtain the same amount of money as he would by establishing a firm, while workers will achieve a level of utility greater than or at least equal to the level they would obtain by working for the monopolist. Under these circumstances, firms are unlikely to be established.

Similarly, firms are not likely to be established in case ii either. Any owner who might start a firm always has the simpler option available, like our monopolist in case i, of renting his land to the workers of his hypothetical firm and charging them a rent equal to the maximum possible profit he could obtain from establishing a firm divided by the number of his employees. Under this arrangement, the owner will receive the same amount of money as he would receive if he set up a firm, while workers will obtain a level of utility which is greater than or at least equal to the level of utility they would obtain by working for the owner as employees.

The results obtained by Calvo and Wellisz in section II of their 1978 paper and our own discussion above lead us to make the following conjecture, which we present as a still unproved theorem: that for any set of conditions under which there is no limit to the size of firms, there is no reason for firms to exist. And this conjecture, together with the fact that firms do exist and are of limited size, in turn suggests to us that the optimal size of firms may be determined by those very conditions that bring about their existence.

References

Calvo, Guillermo A., and Wellisz, Stanislaw. "Supervision, Loss of Control, and the Optimum Size of the Firm." *J.P.E.* 86, no. 5 (October 1978): 943–52.

Williamson, Oliver E. "Hierarchical Control and Optimum Firm Size." *J.P.E.* 75, no. 2 (April 1967): 123–38.

Prior Information and the Observational Equivalence Problem

Warren E. Weber

Virginia Polytechnic Institute and State University

The observational equivalence problem pointed out by Sargent (1976) consists of two parts. The first is whether regression analysis of so-called classical and Keynesian reduced-form equations can discriminate between the competing hypotheses, using data from a single policy regime in the absence of any prior information concerning the policy rule being followed. The answer to this question is no and can be shown as follows.

Consider the reduced-form equation for a "goal" variable (e.g., real GNP). The classical version of this reduced-form equation would be

$$\alpha(L)y_t + \beta(L)(m_t - E_{t-1}m_t) = \mu_t, \tag{1}$$

where m_t is a policy variable (e.g., the nominal money supply), and μ_t is a white-noise error term. This equation is classical, since only innovations in the policy will affect the goal variable. As a consequence, the time path of the goal variable will be independent of the policy rule chosen, so long as the rule implies $m_t = E_{t-1}m_t + \epsilon_t$, where ϵ_t is a white-noise error distributed independently of μ_t.

The Keynesian version of this reduced-form equation would be

$$a(L)y_t + b(L)m_t = \nu_t. \tag{2}$$

Assuming that (2) is invariant with respect to interventions that change the m_t process, it is broadly "Keynesian," since it would be possible to solve (2) for a policy feedback rule which would minimize the deviations of the goal variable around some desired level.

Professor of economics, Virginia Polytechnic Institute and State University. I am indebted to Anatoli Kuprianov, Bennett McCallum, and an anonymous referee for useful comments on this paper.

[*Journal of Political Economy*, 1981, vol. 89, no. 2]

To show that estimation of (1) and (2) will not distinguish between the Keynesian and classical positions, suppose that the policy variable follows the rule

$$\gamma(L)m_t = \epsilon_t \tag{3}$$

over the entire observation period. Then substituting (3) into (1), we obtain

$$\alpha(L)y_t + \beta(L)\gamma(L)m_t = \mu_t. \tag{4}$$

In the absence of any prior information on the order of the polynomials $\gamma(L)$ or $\beta(L)$, it would be impossible to distinguish empirically between (4) and (2). That is, if both equations included the same number of lagged terms, they would have the same sum of squared residuals, since $\beta(L)\gamma(L)$ would be an arbitrary polynomial in L. Further, since (4) has an identical error to (1), the sum of squared residuals obtained from estimating (1) would be the same as that from estimating (2), regardless of whether the Keynesian or classical hypothesis were true. Thus, in the absence of any prior information on the lengths of the distributed lags in the model, it is impossible to empirically distinguish between the Keynesian and classical positions using data from a single policy regime.

The second question is whether such a test can be performed when prior information on the order of $\gamma(L)$ or $\beta(L)$ is available. McCallum (1979) has pointed out one such case. Our purpose is to generalize McCallum's result and to relate it to Sargent's original paper and the empirical tests of the classical and Keynesian hypotheses performed by Barro (1977, 1978).

Let $c(L)$ be the polynomial in the lag operator L of the regression coefficients on the m_t terms when equation (4) is estimated; that is, $c(L) = \beta(L)\gamma(L)$. Thus, (4) can be empirically distinguished from (2), if and only if there exist testable restrictions on the relationship between $\gamma(L)$ and $c(L)$. To examine the nature of these restrictions, first note that a typical element of the convolution $c(L) = \beta(L)\gamma(L)$ is

$$c_j = \sum_{k=-\infty}^{\infty} \beta_k \gamma_{j-k}.$$

Thus, $c(L) = \beta(L)\gamma(L)$ can be written in matrix form as

$$C = H(\gamma)B, \tag{5}$$

where $H(\gamma)$ is the $(n + m + 1) \times (n + 1)$ matrix

$$H(\gamma) = \begin{bmatrix} \gamma_0 & 0 & 0 & \cdots & 0 & 0 \\ \gamma_1 & \gamma_0 & 0 & \cdots & 0 & 0 \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma_n & \gamma_{n-1} & \cdot & \cdots & \gamma_1 & \gamma_0 \\ \gamma_{n+1} & \gamma_n & \cdot & \cdots & \gamma_2 & \gamma_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \gamma_m \end{bmatrix}, \tag{6}$$

$B = (\beta_0, \beta_1, \ldots, \beta_n)'$ is the vector of coefficients of $\beta(L)$, and $C = (c_0, c_1, \ldots, c_{n+m})'$ is the vector of coefficients of $c(L)$. Then by the result of Rothenberg (1973, p. 37), knowledge of m or n will yield restrictions relating the coefficients of $\gamma(L)$ and $c(L)$ if rank $[H(\gamma)] < (n + m + 1)$.

With this result, we can now interpret Sargent's original argument and McCallum's result. Sargent considers the case in which $n = m = \infty$. In this case, the restriction that the rank of $H(\gamma)$ be less than $n + m + 1$ is not satisfied, and no restrictions between $\gamma(L)$ and $c(L)$ are obtained. McCallum considers the case in which $n = 0$, $m = \infty$. In this case rank $[H(\gamma)] = 1 < \infty$, and the prior information given by knowledge of $\gamma(L)$ implies restrictions on $c(L)$.

Rothenberg's result, however, allows a more general statement of when knowledge of n or m implies that restrictions between $\gamma(L)$ and $c(L)$ exist. This general result is that knowledge of n or m will imply that testable restrictions between $\gamma(L)$ and $c(L)$ exist, unless $n = \infty$ or n is finite and $m = 0$. This generalization follows directly from the rank condition stated above. It implies that unless y_t depends on all past innovations in policy or policy is white noise, the classical reduced-form equation (4) can be tested against the Keynesian reduced-form equation (2). The test is whether imposing the restrictions of equation (5) leads to a statistically significant decrease in the explanatory power of the regression equation.

The test of the classical model by Barro used the policy equation

$$\gamma(L)m_t + \lambda(L)z_t = \epsilon_t \tag{3'}$$

in place of (3), where z_t is an exogenous variable (such as the minimum wage or level of unemployment compensation) and $\lambda(L)$ is a polynomial of order r in L with $\lambda_0 = 0$. Substituting (3′) into (1) would yield the new classical reduced form

$$\alpha(L)y_t + \beta(L)\gamma(L)m_t + \beta(L)\lambda(L)z_t = \mu_t \tag{4'}$$

to be compared with the appropriately modified Keynesian version

$$a(L)y_t + b(L)m_t + d(L)z_t = \nu_t. \tag{2'}$$

Now let $f(L)$ be the polynomial in L of the regression coefficients on the z_t term when (4′) is estimated, so that $f(L) = \beta(L)\lambda(L)$. Then using the same arguments as above, (4′) can be empirically distinguished from (2′) if there exist testable restrictions between $\gamma(L)$ and $c(L)$ or between $\lambda(L)$ and $f(L)$. The case of $c(L)$ was considered before. Following the reasoning used for that case, we find that knowledge of r will yield testable restrictions between $\lambda(L)$ and $f(L)$ if and only if rank $[H(\lambda)] < (n + r + 1)$. Since $H(\lambda)$ is $(n + r + 1) \times (n + 1)$, knowledge of r will not imply that testable restrictions between $\lambda(L)$ and $f(L)$ exist when $n = \infty$. Thus, Barro's method of including exogenous variables in the policy rule which do not appear in (1) will not yield testable restrictions to distinguish classical and Keynesian reduced forms when y_t depends on all past innovations in the policy. However, when n is finite and $m = 0$, so that no testable restrictions between $\gamma(L)$ and $c(L)$ exist, the knowledge that $r \geqslant 1$ implies that testable restrictions between $\lambda(L)$ and $f(L)$ exist. Thus, in this case, Barro's method will permit the classical and Keynesian reduced-form equations to be distinguished when they could not be if z_t were omitted from the policy rule.

The practical usefulness of these results, of course, depends on the existence of reliable theoretical restrictions on the order of the lag distributions of $\beta(L)$, $\gamma(L)$, or $\lambda(L)$. The order of $\beta(L)$ depends on the extent to which past errors in the expectation of policy affect the current level of the goal variable. One source of restrictions on the value of n can be found in the models of aggregate supply developed by Lucas (1972, 1973) and Leiderman (1979) in which producers in informationally separated markets face the problem of distinguishing relative price changes from absolute price changes, given current information only on the prices they receive. In these models, output is affected only by the current error in the expectations of the price level implying $n = 0$. The same restriction on the value of n arises from Sargent's (1978; 1979, chap. 16) models of employment. McCallum also argues in favor of such a restriction on the grounds that "it is hard to imagine ways in which past expectational errors could have direct effects on current behavior—bygones are, after all, bygones" (1979, p. 398). Thus, there appears to be a solid theoretical basis for imposing the restriction that $n = 0$ on $\beta(L)$.[1]

[1] This conclusion is not altered by the consideration of adjustment or search costs. Such costs may lead to past expectational errors having indirect effects on the current

Given $n = 0$, the rank conditions above indicate that testable restrictions between $c(L)$ and $\gamma(L)$ or between $f(L)$ and $\lambda(L)$ will exist unless policy is white noise (i.e., unless $m = r = 0$). As yet, there is no satisfactory theory of how policy functions are determined, which would allow us to theoretically determine that policy is not white noise. However, the evidence from empirical attempts to estimate monetary policy functions does support such a conclusion.[2]

Thus, we would argue that reliable theoretical and empirical evidence exists to allow testing of models with classical implications against models with Keynesian implications, using data from a single policy regime. However, the lack of theoretical restrictions on m or r does caution that such tests should check for their robustness to changes in the length of the lags assumed in the policy functions underlying the test.

References

Barro, Robert J. "Unanticipated Money Growth and Unemployment in the United States." *A.E.R.* 67 (March 1977): 101–15.

———. "Unanticipated Money, Output, and the Price Level in the United States." *J.P.E.* 86, no. 4 (August 1978): 549–80.

Froyen, Richard T. "A Test of the Endogeneity of Monetary Policy." *J. Econometrics* 2 (July 1974): 175–88.

Leiderman, Leonardo. "Expectations and Output-Inflation Tradeoffs in a Fixed-Exchange-Rate Economy." *J.P.E.* 87, no. 6 (December 1979): 1285–1306.

Lucas, Robert E., Jr. "Expectations and the Neutrality of Money." *J. Econ. Theory* 4 (April 1972): 103–24.

———. "Some International Evidence on Output-Inflation Tradeoffs." *A.E.R.* 63 (June 1973): 326–34.

McCallum, Bennett T. "On the Observational Inequivalence of Classical and Keynesian Models." *J.P.E.* 87, no. 2 (April 1979): 395–402.

Neftci, Salih, and Sargent, Thomas J. "A Little Bit of Evidence on the Natural Rate Hypothesis from the U.S." *J. Monetary Econ.* 4 (April 1978): 315–19.

Rothenberg, Thomas J. *Efficient Estimation with A Priori Information.* New Haven, Conn.: Yale Univ. Press, 1973.

Sargent, Thomas J. "The Observational Equivalence of Natural and Unnatural Rate Theories of Macroeconomics." *J.P.E.* 84, no. 3 (June 1976): 631–40.

———. "Estimation of Dynamic Labor Demand Schedules under Rational Expectations." *J.P.E.* 86, no. 6 (December 1978): 1009–44.

———. *Macroeconomic Theory.* New York: Academic Press, 1979.

---

value of the goal variable, because they would affect its value in previous periods. However, such effects would be captured in $\alpha(L)$ (see McCallum 1979, p. 398).

[2] See Froyen (1974), Neftci and Sargent (1978), and the references contained therein for estimates of monetary policy functions which support this conclusion.

Book Reviews

The Enclosure of Ocean Resources: Economics and the Law of the Sea. By Ross D. Eckert.
Stanford, Calif.: Hoover Institution Press, 1979. Pp. 390. $16.95.

This is a clearly written book about the Law of the Sea, applying Coase/Demsetz economic ideas to regulation of, and property in, six or more uses of ocean resources. The central theme is that national "enclosure" by the extension of base lines, and widening of territorial seas and various types of exclusive economic zones, are analogous to the historic "enclosure" of common lands and the appropriation of frequencies and so on. The book closes with chapters on UNCLOS (the United Nations Conference on the Law of the Sea) as a venue in which enclosure is debated and arranged.

Written for a more general audience than academic economists, the book is sometimes less than rigorous in its analysis of rents, profits, and transactions costs and may also be given to sliding over some difficult details of law, geography, engineering, and the other fields in which expertise is needed to write such a book satisfactorily to all. But the documented attempt to put the claims of an economics/law property rights approach is worth pondering.

The author fully realizes that "watertight" exclusivity arrangements for some ocean uses would be astronomically costly or even technically impossible. Consider fisheries, to which he devotes a long chapter. Enclosure of coastal fisheries leads to internalized control over local stocks but is obviously less reliable in controlling the harvesting and growth of migratory species. For the latter, the author faintly recommends (as he does for some types of oil-spill problems) a mixed international-coastal regime. But he warns that, for such arrangements, the transactions costs between nations are high; he conjectures that costs for cooperation, enforcement, and decision making increase seriously with the number of countries participating; and he comes out in favor of single-coastal-nation fisheries' regimes.

In this recommendation he is in distributive deep water, as he recognizes. The alternative to single-nation control is treaty-organized "communal" control, giving some power to all nations. Eckert argues that this system is inferior to that by management by individual states. The evidence of inferiority is only one of cost and effectiveness: that transactions costs do seem high, and success does seem incomplete, in such bodies. On such cost evidence, the author seems inclined simply to argue that single-nation regimes should prevail. But this alternative is not one that is available. Not only are many ocean resource uses unenclosable physically, but many claims to conflicting fishing or other rights are of ancient vintage and cannot be cancelled merely by an economist

[*Journal of Political Economy*, 1981, vol. 89, no. 2]

asserting that fishing success is higher and costs lower when one country takes all.

There *is* an answer, involving multinational equity ownership of the rights to use certain ocean resources, actual management being entrusted by all to one nation or to a private firm. The Pacific fur seal treaty provides an example. Some UNCLOS proposals having to do with minerals and other resources could be interpreted or remolded in this form, but Eckert does not do so. His adamant opposition to large numbers is therefore unconvincing. (A side benefit of this discussion is his analysis of the process of national adjustment to the likely future outcome of UNCLOS bargaining. A country about to make a bargain that depends upon its initial position tries to magnify its initial position. Eckert notes this several times.)

A second answer to the many-nation difficulty is to offer compensation to those who are excluded. If it is too costly to let all nations fish or mine, then Eckert should provide for some way to buy them out. Although the process of doing this would be costly, it might work, and the UNCLOS pursuit of the dogma of the oceans being "the common heritage of mankind" could be utilized here. But Eckert seems to me to give little if any weight to the view that UNCLOS is a constitution-making body from which all must depart better off than they arrived.

In other words, his quickness and clarity in making points against UNCLOS procedures do, it seems to me, blind him to the difficulty that distributional imperatives do call for multination ownership or benefit. Otherwise the injustice of exclusion would match that of the original enclosure movements, which Eckert seems to ignore.

In economists' jargon, Eckert seems strongest on efficiency and weakest in the realms where distribution and public choice meet. This is a pity, for his book ends in frustrated anger at the "hypocrisy" of bargaining at UNCLOS, when his earlier treatment, clean and informative, somewhat extended beyond the domain of operating transactions costs, would suggest the limits of what is possible given the costs of constitution making and the conflict in every nation's breast of distributional and efficiency objectives.

Anthony Scott

University of British Columbia

Issues in Financial Regulation. Edited by Franklin R. Edwards. Regulation of American Business and Industry Series.
New York: McGraw-Hill Book Co., 1979. Pp. 526. $42.50.

This book can best be thought of as an intellectual sampling of approaches to regulatory problems by a diverse group of economists and lawyers. For those who want such a sampling, reading it would be useful. All economists do not think one way about regulation, and all lawyers another. There is a diversity of opinion in both camps. As an economist, I will venture the opinion that the economics profession can offer two criteria that should be applied to analyses of regulatory problems: (1) Are sound theoretical notions applied to the problems? and (2) Is good empirical methodology used? The first criterion involves not merely an understanding of theory but an understanding of the problem, how it arises, and the market forces that surround it. The second criterion involves not merely the gathering and analyzing of information but

asking the right questions about the type, amount, and quality of information needed to analyze the problem. It is useful to look at the papers in the book in this manner. Lawyers will undoubtedly have other criteria, but in my opinion a great number of our regulatory problems have arisen because one of these two criteria has been violated.

The papers in this book indicate that some thinkers pay attention to these criteria, others do not, and the quality of the papers is influenced accordingly. Kenneth Scott analyzes the dual banking system as a model of competition among regulatory agencies. By dual banking system, he means the system of alternative chartering at the state and national level and the system of divided regulation at the federal level that goes along with it. However, he only analyzes one-half of the problem. Competition among regulators lets banks prevent the imposition of unduly burdensome regulation by using the option of a change in chartering status to play one regulator off against another. Scott says this is good. But unless there are other checks on the regulators, the logical end of such a process is no regulation at all—that is, competition in laxity. But other checks do occur. Bank regulation and supervision is a technical business in which the best systems and procedures are far from obvious even to a trained professional. Regulators, like the rest of us, do not like failures, and they learn from each other's mistakes. As in the marketplace, alternatives in bureaucratic regulation promote innovation and efficiency, a form of "competition in excellence." In their periodic oversight hearings on the condition of the banking system, the Congressional banking committees are not just assessing the credibility of the reports of one monolithic regulatory agency. They are comparing the performance of several agencies. It is not obvious that the system produces overregulation or underregulation. But there are inherent checks and balances against either tendency.

Bank regulators have the dual responsibility of preventing extensive bank failures which would bring about unwanted contractions in the money supply and of insuring low-cost delivery of banking services to the public. This dual responsibility should be kept in mind when analyzing proposals for change. Scott presents an extremely complicated organizational chart to depict the structure of the system that regulates commercial banks. The chart displays the interrelationships between the Comptroller of the Currency, the Federal Reserve, the Federal Deposit Insurance Corporation, and the 50 state regulatory agencies. The chart is incomplete. It does not include the Securities and Exchange Commission, the Federal Trade Commission, the Treasury Department, the Department of Housing and Urban Development, the Commerce Department, the Labor Department, and the Energy Department—some of the other agencies that have gotten into the business of regulating banks. In addition, the Federal Home Loan Bank Board, the National Credit Union Administration, and the Farm Credit Administration—all omitted from the chart—are also regulators of deliverers of banking services and interact with the other bank regulators. The chart was constructed by Senator Proxmire's staff for the purpose of justifying consolidation and simplification of federal bank regulation. If we wish to design bank regulatory systems for the administrative convenience of the regulators, the senator's proposed reforms merit consideration. But if the purpose of the regulatory system is to insure continuous and efficient delivery of banking services in the marketplace, the failures of the current system are not obvious.

Failure to apply both criteria is illustrated in Roy Schotland's paper on conflicts of interest. He relies extensively on anecdotal evidence and con-

structed situations with very little analysis of market structure or economic incentives. On page 125 he says that most conflicts occur because competition is not effective, an assertion that is not justified with any evidence. On page 128 he hastens to add that most conflict situations are not exploited. Could it be that competition is effective? He says that confidentiality requirements make "marketplace type shopping" unfeasible for doctors, lawyers, and investment managers. No evidence is offered on the extent to which consumers actually do shop around for the services of such individuals. I make an attempt at it, and I do not think I am that unusual. Schotland makes an issue out of the conflict an investment banker incurs because of his obligation to "the buying public, to corporate issuers, and self-interest." He does not say why this is any different from the problem faced by the president of a manufacturing firm who has to mediate the competing claims of customers, suppliers, and stockholders, or why it should require any more regulation.

The paper by Robert Shay, "Consumer Protection in Credit Markets: An Analysis of Regulatory Reform," is an excellent example of judicious application of both criteria. Shay uses the economic analysis of property rights to assess the regulatory reform movement affecting consumer credit markets. He develops an analytical framework and applies it to one limitation of a creditor's remedy to recover debt: the prohibition of wage assignments. He then compares his analysis of the problem to that used by the Federal Trade Commission (FTC). The comparison is fascinating and revealing. Although he does not do any empirical testing, he develops a precise way of getting answers to the questions needed to implement his theory.

Shay believes a problem in consumer credit markets today is that the focus of bargaining is on rates and terms and not on the remedies and rights of creditors and borrowers. Rather than prohibit wage assignments, parties should be permitted to bargain over whether or not they are to be allowed in the contract. The author notes that when wage assignment provisions are allowed they are not uniformly found in all credit contracts. This suggests that when they are included, the reason is to lower the cost of credit to high-risk applicants.

To assess the need for government interference Shay suggests using Demsetz's principle of compensation. The government should try to buy permission to allow the activity from those who feel they would be harmed by allowing it, and it should try to buy permission to restrict the activity from those who feel they would benefit from allowing it. The lower-cost option should be purchased.

Shay's analysis indicates that the adversary relationship is not between debtor and creditor but between those who will be harmed if there is freedom to contract for remedies and those who will be harmed if there is not. The creditor is best viewed as an intermediary. Only a modest portion of the funds being lent are his own funds. Nevertheless, the creditors' adjustments to governmental intervention determine who gets helped and who gets hurt. The need is for marginal benefit/cost comparison between two groups of debtors, not between the debtors vis-à-vis the creditors. The rather crude balancing process used by the FTC is between the creditors and the debtors.

Two of the papers dealing with electronic funds transfer (EFT) propose radical policy changes without using either of the criteria mentioned above. Almarin Phillips claims that regulation and institutional arrangements—interest rate ceilings, deposit insurance, etc.—cause difficulties in the administration of monetary policy. Undoubtedly this is true. But he then goes on to assert, in language phrased to make the dissenter feel antediluvian, that EFT

makes all of this substantially worse and will bring a crisis. He offers no evidence or analysis to support his conclusion, and I, for one, dissent from it. To be sure, many regulations cause difficulties for private bankers, central bankers, and the public. But it is the regulations that cause the problem, not the use of electronic forms of communication.

Similarly, Alan Westin says that EFT creates the need for new statutes to spell out individuals' rights. He says that when information is merged from a variety of sources and not used for one-time purposes a new trustee relationship is created. Of course, such information gathering is as old as banking itself. But when it is done electronically, Westin believes new laws are called for. He is not exactly sure about what the new laws and regulations should be. He proposes a decade or two of controlled experimental regulation. He gives no discussion of current abuses to support this recommendation.

Arnold Heggestead does a survey of studies of bank competition and performance. Most of the articles surveyed, and the survey itself, suffer from a lack of application of the criteria mentioned here. This is best illustrated by Heggestead's ambiguous and contradictory conclusion. He starts out by stating that local banking markets are "more monopolistic than competitive." He ends by stating five major areas where further empirical research is needed to clear up anomalies.

If he had started out by analyzing the way markets operate he could have explained some of the anomalies himself. Interinstitutional competition is a fundamental fact of today's marketplace. If one examines all of the various asset powers, liability powers, and fee income powers of commercial banks, there is only one privilege that commercial banks have that is not granted to at least one other financial institution. This is the power to offer demand deposits, a privilege whose value is rapidly being eroded by market innovations and changing product mixes. In 1947, commercial banks held 56.3 percent of the assets of all financial institutions. In addition to thrift institutions, this latter category includes life insurance companies, private pension funds, state and local employee retirement funds, money market funds, and finance companies. In 1977 the corresponding figure was 39.6 percent. In 1947, they held 81.7 percent of the total deposits of all depository institutions. In 1977, it was 59.3 percent. In 1947 commercial bank deposits were 56.2 percent of the total liabilities of all financial institutions. In 1977 the corresponding figure was 32.4 percent. Demand deposits, the product for which commercial banks are sometimes said to have a "monopoly," were 40.3 percent of total financial institution liabilities in 1947. In 1977, the corresponding figure was 10.9 percent. This is not the picture of a monopolistic industry. When they had reason to do so, bank customers were highly sophisticated at finding alternative ways of meeting their financial needs.

Some of the other papers are good examples of applications of the criteria set forth here. Others are not. But enough has been said to illustrate the problem.

The book suffers from a lack of discussion of the political aspects of regulation. I refer to the process whereby information is gathered through hearings and testimony, and laws are passed; and more information is gathered through hearings and testimony, and regulations are written. With increasing frequency, outsiders are having a hard time matching the laws with the regulations, and Congress and the regulators are pointing their fingers at each other. I have no final answer to this problem, but I will suggest two partial causes. First, the process inherently favors anecdotal evidence over econometric work and surveys which, because they are subject to

methodological criticisms, inevitably present a more balanced view of the world. Second, when the banking system works well, the bankers get the credit, not the regulators. When it goes awry, the regulators share a good portion of the blame. This inevitably promotes regulations designed to protect inefficient banks, to the detriment of the customers of well-managed banks.

P. Michael Laub

American Bankers Association

New Challenges to the Role of Profit. Edited by Benjamin M. Friedman. Lexington, Mass.: D. C. Heath & Co., Lexington Books, 1978. Pp. viii+127. $14.95.

In 1976, Professors Samuelson, Arrow, and Lundberg presented lectures in the third series of the John Diebold Lectures. Along with an introductory chapter by Professor Friedman and a concluding round-table discussion, these lectures comprise *New Challenges to the Role of Profit.* Samuelson gives a broad-brush treatment of the trend of profits, and intellectual opinion about profits, in the modern mixed economy. He also includes a brief history of profit theory. Arrow links criticism of profits to general attacks on private property, but he emphasizes the conceptual differences between profit, property income, and private property. Lundberg reports on the Swedish "middle way"—now dubbed "fund socialism"—and attempts to draw conclusions for Western countries generally.

All things considered, the book is a disappointment from almost any perspective. Even for a work of this genre, *New Challenges* lacks focus. Though edited by a perspicacious scholar who contributes an introductory essay, the book wanders. The subject remains unclear. The principal lectures were delivered by academics, with comments by a mixture of academics and nonacademics. At a minimum, one would have hoped for careful delineation of a profit concept, distinguishing profit from equilibrium returns to owners of capital, land, and other scarce resources (i.e., interest and economic rent). Though any number of allusions and some specific references are made to the distinction between interest (or ordinary profit), rent, and pure profit, these various incomes are constantly conflated. The analysis is further confused by unwarranted shifting in the lectures between static and dynamic analysis. For instance, in any essay containing a number of fruitful insights, Friedman observes that "real economies are dynamic, not stationary" (p. 4); he then notes the importance of this realization for our understanding of entrepreneurship, innovation, development, and industrial structure. This discussion is immediately followed, however, by a reference to the intertemporal allocation role of profits (more properly an interest problem). Profits are then treated as accruing to owners of capital, thus entangling pure profits, rents, and interest. Shortly thereafter, Friedman considers the trade-off between equity and efficiency—or distribution and allocation—arising in a profit system. Almost every contributor eventually refers to this trade-off. The discussions of the "trade-off" remind one of those concerning the "cruel choice" between inflation and unemployment. There may indeed be different combinations of income distributions and allocation efficiency, depending on the taxes and other constraints placed on profits, but it is by no means clear that any stable social choice set exists from which one could pick and choose

combinations. Further, if one considers the behavior of economies over time, then the question is not whether profits (however defined) contribute to income inequality, but whether the profit *system* produces more or less income equality compared to other, *feasible* economic systems. The latter question is not really addressed in this volume.

Along with Arrow's, Friedman's is one of the better contributions to the book. The shifting from dynamic to static analysis and between profit concepts, even in the introductory chapter, is indicative of problems in the rest of the book. No fruitful analysis of "profit," or of any other concept, is possible without a shared understanding of the nature of the problem being discussed. Such a shared understanding is lacking in *New Challenges.* The contributors could surely have adopted a number of profit concepts or definitions. But regardless of the concept or definition chosen, reasonable consistency in analysis is necessary. In this book there is almost no analysis of profits as the essential signaling and motivating force in a market economy. Despite obligatory references to Schumpeter, no approach treating profit as an income emerging from unexpected changes in economic data is seriously considered or analyzed. The work of Kirzner, identifying genuine profit as a return to superior entrepreneurship—alertness to altered conditions—is totally ignored (Kirzner 1973, 1979). Yet if the related theories of Kirzner, Mises, Schumpeter, and Knight are at all correct, then profit—not some fictitious auctioneer—is responsible for any tendency toward plan coordination and equilibrium in decentralized economies. Coordination is, of course, always incomplete, because change is always occurring. But without appropriable profits, the decentralized market system cannot work.

The Kirznerian profit theory clarifies a number of conceptual problems. Profit, being a return to a special labor skill, is no longer conflated with interest, a return associated with capital accumulation. Nor is profit a rent or quasi rent accruing to a fixed or temporarily fixed factor, as in standard textbook analysis. Since profit disappears as quickly as it is publicly identified, and since it is not an economic rent, profit is not a taxable surplus. Arrow and Lundberg suggest that the ownership of property and the receipt of property income can be separated without serious allocational consequences. As Arrow put it: "This justification of the charging of price for the use of property is not necessarily an argument that the reward should go to the owner rather than to someone else" (p. 53). This line of argument is more plausible if profit is considered as a residual or rentlike income. In Kirzner's analysis, however, profits are the driving force of economic activity. They are actively and consciously strived for. They neither accrue to passive resource owners nor are they a residual. Thus, profit income and private property are inextricably linked and intertwined. As Marxists have always understood, the fundamental issue is the private property system and the very existence of markets.

The latter point helps indicate the deficiency of the book. Most of the challenges to profits discussed are quite old. Yet the contributions do not build on the long literature on the subject (Arrow's essay again is the exception). Newer critiques of profit and the market system derive from "highbrow opinion" and not from popular concerns (to which the book is ostensibly addressed). The ordinary person is not concerned that market economies do not attain global maxima and are never in general equilibrium. (Whoever suggested they would be?) The ordinary citizen is becoming aware, however, of the tendency for governmental policy in a mixed economy to substitute a system of monopoly rents and arbitrary incomes for that of earned profits (see Hughes 1977). He may lack the theoretical background to explain the

process, but the man in the street is surprisingly on the mark in his judgments. Lundberg tells us that Western economies will, like Sweden's, become more mixed (p. 91). If so, monopolization and socialization of earned income will become more widespread.

In the mixed economy, or fund socialism, the criticism that profits do not motivate innovation or lead to plan coordination changes from an erroneous critique to an essentially correct one. Inter alia, as measured profits are taxed at higher rates and supplanted by explicit and implicit subsidies, their motivational and allocational importance is diminished. Yet this process will remain largely hidden if viewed from a perspective in which profit and economic rent cannot be clearly distinguished in an analytical sense. In the latter view, "profit" is a type of surplus that, at least within some range, can be altered with little impact on resource allocation.

If, in fact, this book were to help focus attention on the distinctive operation of a mixed, as opposed to a market, economy, then it would have made a contribution. Were the book to help spark a public debate on this question, its faults could be overlooked. But if such a debate were to occur, it would probably be more in spite of than because of such efforts as this book.

GERALD P. O'DRISCOLL, JR.

New York University

References

Hughes, J. R. T. *The Governmental Habit: Economic Controls from Colonial Times to the Present.* New York: Basic, 1977.

Kirzner, Israel M. *Competition and Entrepreneurship.* Chicago: Univ. Chicago Press, 1973.

———. *Perception, Opportunity, and Profit: Studies in the Theory of Entrepreneurship.* Chicago: Univ. Chicago Press, 1979.

Economic Equality and Fertility in Developing Countries. By ROBERT REPETTO. Baltimore: Johns Hopkins University Press, 1979. Pp. 206. $14.50.

This work is best viewed as a collection of pieces in which a model of the relationship between income and fertility is developed, applied, and in turn tested through these applications. The questions raised by the author are crucial ones for population policy, his answers are well reasoned, and if there is still room for disagreement regarding his policy conclusions it is nevertheless much to his credit that the evidence and the issues are well laid out in the book itself.

The book opens with a statement of its central theme: that fertility is a nonlinear negative function of income such that the marginal effects of an income change are greater for poor people (although perhaps not for the abject poor) than for the rich. It follows from this nonlinearity that a given change in income level would have little impact on fertility for families in the middle and upper end of the income distribution but might very well have a large impact on fertility for poor families. The policy implication is that redistribution of income from rich to poor would generate a decline in aggregate fertility and hence population growth.

Having stated this hypothesis in chapter 1, chapter 2 develops a cross-country model for testing its implications. Here it is necessary to assume that the basic structural relationship between fertility and income at the family level is essentially the same in every country (ceteris paribus). If so, it follows quite easily that any redistribution of income from the relatively rich to the poor, including transfers from rich countries to poor countries, would reduce aggregate fertility. It also follows that the relationship between aggregate fertility and income is not fully specified by the level of national income without some measure of its distribution. Chapters 3–5 are country studies establishing the robustness of the basic relationship. Chapter 6 concludes the book with an analysis of worldwide fertility along the lines suggested above.

Each of the country studies uses some multivariate analysis based on household data as well as more aggregated descriptive statistics. In general, the regressions are less satisfying than the descriptive analyses. In part this may be because of difficulties obtaining data for the appropriate concepts and interpreting results based on imperfect data. Largely, however, it is because the separate country studies have not been well integrated into a coherent whole. Although the relationship between income and fertility is the only point at issue, a fairly complex country-specific model is needed in order to isolate that relationship. Despite the claim that chapter 2 develops a general model, each chapter contains its own model which is not clearly related to the "general" one. Moreover, the development of these models, the discussions of the data, and the interpretation of results are all too cursory to be satisfying, even though the limited purpose of this study may not warrant a full country analysis for each case. The author's analysis is much more convincing when he is using his model to tell us a story about country experiences. We are thus left with the impression that he may know more than he has revealed in his country studies.

The case study of Puerto Rico has relatively little descriptive analysis; it uses 1970 census data to estimate the magnitude of an income effect on fertility for households at different income levels and to provide empirical confirmation of the nonlinear relationship. The second case study shows how such a relationship can be used to interpret the Korean experience: a major redistribution of income followed by both growing income and declining fertility. This is the most thorough of the case studies, and it is perhaps no coincidence that the regression model is limited to a few pages at the end.

By far the most disturbing part of this book is the chapter entitled "Internal Policies for Income Distribution in Rural India." The data from Matar Taluka (a rural area in India's Gujarat state) are fascinating in that they resurvey an area for which data were also collected in 1930 and so permit an analysis of long-run changes in income distribution. These data are used as a point of departure for an analysis of the benefits of land reform; fertility comes up only once, in passing, when the author asserts that redistribution through land reform will have the desired income effect. The discussion of land reform which takes up fully half of this chapter is well presented and argued, but its relevance to the main theme or to anything else in this book is established tenuously at best. First, an income regression equation is estimated for Matar Taluka in which both human and nonhuman assets are included as independent variables, and it is found that variation in the nonhuman asset variables accounts for by far the largest portion of the explained variation in income. However, the author tells us that Matar Taluka is an agricultural region with large variation in the size of landholdings and that the survey year was a disastrous agricultural year for most of India;

ceteris paribus, this should lead us to expect a wider-than-usual distribution of income correlated with each farmer's harvest within the region. At the same time, we are told why the human capital variables will not predict income: because there is no variation in schooling within the region and because age pertains to the household head instead of the individual earner. Thus the empirical results for Matar Taluka may tell us what was going on during the survey year but certainly cannot justify choosing land reform over human capital development as a long-term strategy.

Repetto follows his analysis of the effects of land reform with some arguments why education cannot be effectively used to redistribute income through human capital stocks, even though that is in fact what he says has happened in Korea. His arguments often confuse inequality measures with the distribution of income. For example, since the initial distribution of schooling in Matar Taluka is very equal, the first-round impact of increasing the education level of younger cohorts will be to increase inequality of earnings, presumably *because* education improves earnings. Even ignoring potential intergenerational transfers, this is surely a most extraordinary argument against progress, and one which I doubt would impress many low-income families. In the same vein, the following paragraph contains the suggestion that since improved education would affect only new entrants to the labor force, it would be decades before their numbers could be great enough to appreciably affect the average income. Then, having observed that an important value of education is that it "prepares students . . . for jobs outside the traditional sector," he says that land reform is needed because education "has more value for a future landholder than for a future agricultural laborer" (p. 149).

The irony of this analysis of the relative merits of human and nonhuman assets as sources of increased income for the poor is that although it is a small part of Repetto's book it is strategically placed at a point where policy implications are being drawn from research findings. The fact that the discussion of this point is at best superficial is especially puzzling in the light of the comparatively careful reporting of evidence in the earlier chapters. Indeed, on the whole this book contains a useful perspective on the importance of income effects on fertility and provides convincing evidence that the improvement of incomes for the poor, especially in developing countries, is necessary and even perhaps sufficient as a condition for reduced aggregate fertility.

CARMEL U. CHISWICK

University of Illinois at Chicago Circle

The Economics of the Performing Arts. By C. D. THROSBY and G. A. WITHERS. New York: St. Martin's Press, 1979. Pp. 348. $27.50.

Books on the economics of the performing arts are no longer a rarity. But this volume contains material quite different from that of its predecessors. It goes over and brings up-to-date the statistics on costs, revenues, and public support which have been analyzed before. But, in addition, it seeks to go as far as the subject allows in applying the standard techniques economists have used to analyze other industries—adapting theoretical micro models and econometric analysis to the problems of the arts.

Predictably, the success of this venture is mixed. But this is no discredit to

the authors, who are obviously masters of both the subject area and the techniques, and who display consistent good judgment on issues of public policy, which is the book's ultimate focus.

The volume deals with all of the subjects which one would expect it to consider, starting with the organizations which present performances ("the firms"), going on to consumer demand, industry structure, and then turning to public assistance to the arts.

Most subjects are treated in four different ways. Each is discussed in general terms to describe the most pressing issues to which it gives rise. Second, extensive and often illuminating descriptive statistics are provided. Third, where possible, a formal theoretical model is constructed. Finally, an econometric model is formulated and the values of its parameters are estimated.

One can quibble here and there with the more discursive portions of the book, but, by and large, they are accurate and illuminating. Of particular value is the extensive material comparing performing activity and its financing in Australia, Canada, Great Britain, New Zealand, and the United States. The value of such international comparisons has often been asserted, but in the past the materials provided for the purpose have generally been quite spotty. The major conclusions that emerge from the comparisons should, however, hardly be surprising to anyone with some acquaintance with the field. They show remarkable similarity in the economic problems faced by the arts in the different countries, in the characteristics of the audiences, in the nature of the organizations, etc. For example, it is "striking" that "proportions of the population exposed to the performing arts are substantially higher for the middle-aged, high income, high education, professional, managerial and white-collar groups" (p. 96). About the only major difference seems to be the heavy reliance on private foundations and individuals for financial support in the United States, while elsewhere such income flows almost exclusively from government. Far more remarkable are the figures on the growth of government support in the five countries studied. For the 8 years 1969 through 1976, in every one of the countries government assistance grew faster than price level, population, average hourly earnings of labor, GDP, and (with the exception of Canada) aggregate government expenditure. In Australia and the United States the degree by which government outlays on the arts outpaced these other indices was particularly remarkable. All of this is noteworthy because in the past periods of inflation have been times of financial difficulty for the arts. As a matter of fact, more recent data for the United States indicate that government outlays on the arts have continued to rise in real terms, so that while the arts have experienced financial difficulties stemming from other sources, at least so far, and with some noteworthy exceptions, real cuts in government funding have *not* been among them.

The authors' discussion of policy deals with such issues as justifications of assistance, the determination of the appropriate assistance levels and of the allocation of the funds by size of organization, art form, geographic location, etc. In discussing justifications for support the authors build their conclusions on the economists' usual grounds for government intervention: externalities, redistribution, merit goods, option demand, etc. The discussion strikes one as sensible, though no surprising conclusions emerge. They reconfirm what economists had noted before—how difficult it is to provide a very strong justification for government support of the arts in these terms except perhaps on the merit goods argument. But, then, the term "merit good" merely becomes a formal designation for the unadorned value judgment that the arts

are good for society and therefore deserve financial support. In other words, the merit good approach is not really a justification for support—it merely invents a bit of terminology to designate the desire to do so.

As already noted, the unique contribution of this book lies in its theoretical and econometric models, and these are, therefore, worth considering in a bit more detail. These models deal with a variety of subjects: consumer demand for leisure (following Buristam Linder); willingness to offer philanthropy; the theory of merit goods; propensity to provide philanthropy; and the behavior of the arts firm and its response to subsidy. The last of these is perhaps representative. Taking the nonprofit arts firms to be managed by utility maximizers constrained by a zero-loss condition, the authors proceed to formulate a utility function whose variables are attendance and quality, a production function which takes attendance to be a function of number of performances, number of distinct productions, capacity of auditorium, and quality. This is related to a cost function, whose variables are attendance and quality, and a demand function dependent on price and quality. From this, a standard partial-equilibrium diagram emerges, showing the equilibrium of the utility-maximizing firm and its relation to that of a profit maximizer. The conclusions which emerge are not particularly surprising—the utility maximizer's output (audience size) will be the larger, and its output will be increased by subsidy—even if the subsidy is a lump sum.

The econometric models also cover a wide variety of subjects: theater cost functions, consumer demand functions, voter characteristics, magnitude of government assistance, and the effects of grants on attendance. The relatively simple forms taken by the relationships used in estimation are fully explained by the complexity of the subjects and scarcity of data. More difficult to explain is the rather persistent failure of the authors to provide interesting and testable hypotheses. It is as though the statistical calculations were undertaken as ends in themselves rather than as means to get at more fundamental issues. Nevertheless, some very interesting results emerge. For example, the study of the effects of grants indicates:

> Firstly, grants to drama have been more effective at the margin in securing greater attendances than have grants to opera, ballet and music. Secondly, the marginal unit of subsidy in drama has had a relatively higher effect on attendances at low levels of grants per company, declining as the levels of subsidy increased, although this diminishing productivity at the margin is not so evident in the case of the other art forms.
>
> It must be remembered that these results obtain at the margin in a situation where the existence of grants already accounts for a major part of attendances, especially in the case of opera, ballet and music. . . . In the case of theatre, however, the wider range of levels of activity and the more profound effects of a relatively small amount of support on a small company give greater scope for grants to lead to increased attendances in the lower ranges, although at the other end of the scale (i.e., for the major subsidized companies) the marginal unit of grants is less effective in increasing attendances for reasons similar to those discussed above in relation to opera, ballet and music. [Pp. 268–69]

To summarize, this book is a very useful addition to the literature. It will prove indispensable to anyone conducting research on the subject. The au-

thors have given us up-to-date information, have contributed valuable insights, and have expanded the armory of weapons shown to be helpful. But their attempt to fit the analysis to the procrustean bed of standard economic techniques seems to have been pushed beyond the point of diminishing returns.

William J. Baumol

Princeton and New York Universities

Hilda Baumol

Mathematica, Inc.

Journal of Political Economy

Volume 89, Number 3, June 1981

The University of Chicago Press

JOURNAL OF
POLITICAL ECONOMY

Edited by

JACOB A. FRENKEL SAM PELTZMAN
ROBERT E. LUCAS, JR. GEORGE J. STIGLER

In cooperation with OTHER MEMBERS of the DEPARTMENT OF ECONOMICS and the GRADUATE SCHOOL OF BUSINESS of the UNIVERSITY OF CHICAGO and OUTSIDE REFEREES

Editorial Assistants: VICKY M. LONGAWA and LISE A. PLOTKIN

The Journal of Political Economy (ISSN 0022-3808) is published bimonthly in February, April, June, August, October, and December by the University of Chicago Press. Subscription rates, U.S.A.: institutions, 1 year \$30.00, 2 years \$54.00, 3 years \$76.50; individuals, 1 year \$22.00. 2 years \$39.60, 3 years \$56.10. Student subscription rate, U.S.A.: 1 year \$16.00 (letter from professor must accompany subscription). Other countries add \$2.50 for each year's subscription to cover postage. Single copy rates: institutions \$5.00, individuals \$4.00. Back issues are available from 1962 (vol. 70). Make all remittances payable to *Journal of Political Economy*, The University of Chicago Press, in United States currency or its equivalent. **Business correspondence** should be addressed to The University of Chicago Press, 5801 Ellis Avenue, Chicago, Illinois 60637.

Claims for missing numbers should be made within the month following the regular month of publication. The publishers expect to supply missing numbers free only when losses have been sustained in transit and when the reserve stock will permit.

Letters to the editors and manuscripts should be addressed to the Editor of the *Journal of Political Economy*, 1126 East 59th Street, Chicago, Illinois 60637. **Manuscripts should be submitted in triplicate, accompanied by a \$40.00 submission fee made payable to the Journal.** The proceeds from the submission fees are used to pay for refereeing services. Accepted manuscripts must be typed according to the University of Chicago *Manual of Style*. References should be typed double-spaced at the end of the article. Footnotes should be numbered in sequence and double-spaced following the references. Tables should follow the footnotes. Originals of the figures, drawn in india ink, should be submitted if the manuscript is accepted. Abstracts not exceeding 100 words should be submitted in duplicate along with the manuscript.

Reprinted volumes 1–72 available from Walter J. Johnson, Inc., 355 Chestnut Street, Norwood, New Jersey 07648. Volumes available **in microfilm** from University Microfilms, 300 North Zeeb Road, Ann Arbor, Michigan 48106; **in microfiche** from Johnson Associates, P.O. Box 1017, Greenwich, Connecticut 06830.

Notice to subscribers: If you change your address, please notify us and your local postmaster immediately, giving *both* your old and your new address. *Allow four weeks for the change.* **Postmaster:** Send address changes to *Journal of Political Economy*, 5801 Ellis Avenue, Chicago, Illinois 60637.

Second-class postage paid at Chicago, Illinois, and at additional mailing office.

Journal of Political Economy

Volume 89 Number 3 June 1981

Articles

Review Article

Comments

The Deterrent Effect of Antitrust Enforcement

Michael Kent Block and Frederick Carl Nold

Hoover Institution

Joseph Gregory Sidak

Stanford Law School

In this paper we formulate and test a model of collusive pricing in the presence of antitrust enforcement. We show that a cartel's optimal price is likely to be neither the competitive price nor the price that the cartel would set in the absence of antitrust enforcement but rather an intermediate price that depends on the levels of antitrust enforcement efforts and penalties. Our empirical results reveal that increasing antitrust enforcement in the presence of a credible threat of large damage awards has the deterrent effect of reducing markups in the bread industry.

Soon after the passage of the Sherman Act, the Supreme Court determined that horizontal minimum price fixing was so inherently injurious to consumer welfare that it should be illegal per se. Horizontal collusion has since become a major focus of federal antitrust

We wish to thank Cayetano Paderanga and Timothy Moore for valuable research assistance, and we gratefully acknowledge the helpful comments of Donald Baker, William Baxter, Gary Becker, Timothy Bresnahan, Aaron Director, George Hay, Peter Pashigian, Sam Peltzman, Mitchell Polinsky, Robert Reynolds, Kenneth Scott, George Stigler, Robert Stillman, members of the Industrial Organization Workshop at the University of Chicago, and participants in seminars at Stanford Law School, UCLA, Cornell Law School, the Rand Corporation, and the Antitrust Division, U.S. Department of Justice. This paper was prepared under grant nos. 75-NI-99-0123, 77-NI-99-0071, and 79-NI-AX-0071 from the National Institute of Law Enforcement and Criminal Justice, Law Enforcement Assistance Administration, U.S. Department of Justice. The opinions stated in this paper do not necessarily represent those of the U.S. Department of Justice.

[*Journal of Political Economy*, 1981, vol. 89, no. 3]

enforcement.[1] Through the use of criminal and civil sanctions the Department of Justice (DOJ) has sought not only to remedy specific instances of price fixing, but also to achieve general deterrence of potential price fixing.

This paper is the first systematic attempt to estimate the impact of antitrust enforcement on horizontal minimum price fixing. We develop a simple theoretical model of the collusive pricing decision and then, using data on the bread industry, assess empirically the deterrent effect of public and private antitrust enforcement on the decision to collude.

I. A Simple Model of Collusive Pricing in the Presence of Antitrust Enforcement

A. Definitions and Assumptions

We construct a relatively simple model to consider explicitly the effect of antitrust enforcement on the decision of firms within an industry to fix prices collusively. The model uses the following variables and definitions:

$p =$ price of output;

$Q \equiv Q(p) =$ demand for the industry's output;

$C \equiv C(Q) =$ total cost of industry output, including a normal rate of return;

$c \equiv C(Q)/Q =$ average cost;

$mc \equiv \partial C/\partial Q =$ marginal cost;

$\lambda \equiv (p - mc)/mc =$ markup over marginal cost;

$\pi =$ total profits of colluders;

$\gamma =$ level of enforcement efforts directed toward detecting horizontal collusion;

$F \equiv t(p - mc)Q =$ the combined civil and criminal penalty for price fixing, where t is the anticipated damage multiple;

$d \equiv d(\lambda, \gamma) =$ the probability that a collusive pricing scheme will be detected.

The definitions of F and d are essential to our analysis and warrant further discussion.

Our specification of the penalty function, $F \equiv t(p - mc)Q$, reflects that under current statutes a price-fixing conspiracy is subject to both criminal and civil sanctions.[2] The Clayton Act's private treble dam-

[1] Posner (1970, p. 398, table 23) reports that 989 of the 1,551 Department of Justice antitrust cases between 1890 and 1969 contained charges of horizontal conspiracy.

[2] The maximum criminal sanctions for price fixing are imprisonment for 3 years, a fine of $100,000 per individual, and a fine of $1 million per corporation. We do not consider injunctive remedies in this model.

age remedy is the most formidable civil sanction and clearly relates to markups, as our specification of F reflects. If criminal sanctions are small—as they usually are—then civil sanctions provide most of the sting of antitrust enforcement, and our specification of F accurately describes the true penalty for price fixing.[3] Moreover, if, when sentencing price fixers, a district court judge considers the size of the markup to be an index of the cartel's perniciousness, then F is still reasonably descriptive on the rare occasions when criminal fines and sentences are significant relative to private treble damage awards.

In the presence of uncertain antitrust enforcement the decision to collude depends not only on the magnitude of the penalty but also on the probability of detection, which we assume increases with the markup, $[\partial d(\lambda, \gamma)]/(\partial\gamma) > 0$. The closer colluders come to the price that they would charge if no antitrust enforcement existed, the greater their chances of being discovered. Two observations support this hypothesis. First and most obvious, higher markups make customer complaints to the DOJ more likely.[4] Second, as Stigler (1968, pp. 268–70) has noted, a positive relationship probably exists between a conspiracy's detectability and its ability to prevent cheating by individual colluders. For example, a joint-sales agency would assure a set of colluders strict adherence to a monopoly price, but so visible a collusive device would be nearly assured of detection. In general, we hypothesize that the more efficiently a collusive device produces markups, the more likely it will be detected and the colluders convicted.[5] Hence, as we have assumed, the very technology of collusion makes it likely that the probability of detection increases with the markup.[6]

We also assume that the DOJ never charges noncolluding firms:

[3] A comparison of the magnitude of criminal and civil penalties appears in Sec. II. Though a substantial increase in potential criminal sanctions was legislated late in 1974 under the Antitrust Procedures and Penalties Act, there is little evidence of its effectiveness in increasing realized sanctions (see Burnham 1978).

[4] Hay and Kelley (1974) find that complaints by customers are the third most numerous method of detection among DOJ price-fixing cases.

[5] Stigler (1968) presents some indirect evidence on this point. In a sample of DOJ cases he divides the types of collusion between what he considers *efficient* and *inefficient* collusion. The average time from inception of alleged collusion to complaint is significantly shorter for the efficient forms of collusion.

[6] Our study of the bread industry revealed, as tentative empirical evidence of the relationship between the probability of detection and the markup level, that the probability of being investigated for price fixing was positively related to the markup level. We assumed that a lag exists between the price-fixing decision and its detection; then, using a one-period lagged value of our markup indicator as an explanatory variable, we estimated a logit model of the probability that DOJ would initiate an investigation for bread price fixing. The estimated intercept for our 208-observation sample was -2.49, and the estimated coefficient on the lagged-markup indicator was 5.69 with a standard error of 2.43. When the logit function is evaluated using the sample mean values, this coefficient implies that a 1-percentage-point increase in the markup increases the probability of an investigation by 5 percent. Sec. II elaborates on our specification of the markup and the nature of our sample.

$d(0, \gamma) = 0$. We assume further that DOJ does not necessarily detect a pure cartel price with certainty—that is, $d(\lambda^c, \gamma) < 1$, where λ^c is the markup that colluders would choose if no antitrust enforcement existed.

B. *Antitrust Enforcement and Optimal Collusion*

The objective of collusion is to set a price that maximizes expected cartel profits.[7] To model this price-fixing decision simply, we impose three additional assumptions. First, firms produce output under conditions of constant marginal and average cost so that $mc = c$. Second, all firms take part in the joint-pricing decision. Third, the most significant cost of any collusive device is its impact on the probability of detection. For example, the major cost of a joint-sales agency is its visibility, not its resource costs. This third assumption implies that, in the absence of antitrust enforcement, firms could duplicate perfectly the price and output levels of a monopoly.[8]

Formally, the colluders set the price p by maximizing the expected value of their objective function, $E\pi$. That is, the colluders

$$\max_p E\pi = (1 - d)\pi_1 + d\pi_2 \equiv Z, \tag{1}$$

where $\pi_1 \equiv (p - c)Q(p)$ is the profit level if the colluders avoid detection, $\pi_2 \equiv [p - c(1 + \lambda t)]Q(p)$ is the profit level if the colluders are detected, and $\lambda = (p - c)/c$ is the markup under constant costs. The necessary condition for an internal maximum is

$$Z_p = -\frac{\partial d}{\partial \lambda}\lambda t Q(p) + [Q(p) + (p - c)Q'(p)](1 - dt) = 0. \tag{2}$$

Equation (2) has a simple yet intriguing implication: If, as we have assumed, the markup significantly affects the probability of detection, then neither the competitive price ($\lambda = 0$) nor the pure cartel price markup ($\lambda = \lambda^c$) can satisfy equation (2).[9] Antitrust penalties, therefore, do not necessarily eliminate price fixing. They are, however, likely to reduce the optimal markup.[10]

[7] This assumes that the colluders are risk neutral. An analysis that allows for risk aversion is presented in Block, Nold, and Sidak (1978). The results of that analysis do not differ substantially from those presented here.

[8] As long as we restrict our attention to industries with a history of conspiracy, considering intraindustry cheating would complicate the model without significantly adding to our understanding of collusive pricing in the presence of antitrust enforcement.

[9] The pure cartel markup is the markup λ^c satisfying the condition $Q(p) + (p - c)Q'(p) = 0$.

[10] If the detection probability were not sensitive to the markup level, then the decision to price fix would be a simple either/or proposition. In this case, as long as $(1 - dt) > 0$, the colluders would set the price at the pure cartel level. Clearly no price fixing would

As long as price fixing is a favorable gamble for some markup, $(1 - dt) > 0$—and our assumptions assure it will be—then risk-neutral colluders obviously will not choose the competitive price. Also, although the "unfairness" of the price-fixing gamble at the pure cartel markup, $[1 - d(\lambda^c, \gamma)t] < 0$, is sufficient to dissuade the colluders from choosing $\lambda = \lambda^c$, it is not necessary. In general, colluders will stop marking up when both $[Q(p) + (p - c)Q'(p)]$ and $(1 - dt)$ are positive. To deter colluders from choosing the full cartel markup is rather easy; to deter all collusive pricing in a conspiracy-prone industry is nearly impossible.

C. *The Effects of Changes in Enforcement Efforts and Penalties on Collusive Markups*

To assess the effect on markups of increases in enforcement efforts, we consider the effect of changes in γ on the optimal λ, or $\partial\lambda/\partial\gamma$. Totally differentiating equation (2) with respect to λ, we obtain

$$\frac{\partial \lambda}{\partial \gamma} = \left(\frac{t}{cZ_{pp}}\right)\left\{\frac{\partial^2 d}{\partial\lambda\partial\gamma}\lambda Q + \frac{\partial d}{\partial \gamma}[Q(p) + (p - c)Q'(p)]\right\}, \qquad (3)$$

where Z_{pp} is the second-order derivative. Only in the rather perverse case where an increase in enforcement actually reduces the impact of the markup on the detection probability—that is, where $(\partial^2 d)/(\partial\lambda\partial\gamma) < 0$—does an increase in enforcement have an ambiguous effect on the markup. In most cases of practical importance we expect that DOJ efforts designed to increase the probability of detection also reduce the markup.

Increasing price-fixing penalties is often advocated as an efficient means to achieve deterrence.[11] The effect of such an increase is expressed by the comparative static derivative:

$$\frac{\partial \lambda}{\partial t} = \left(\frac{1}{cZ_{pp}}\right)\left\{\frac{\partial d}{\partial \lambda}\lambda Q + d[Q(p) + (p - c)Q'(p)]\right\}, \qquad (4)$$

which indicates that a higher penalty unambiguously decreases the optimal markup. In summary, an increase in either the probability of

occur if $(1 - dt) < 0$. Enforcement efforts would determine simply whether or not it paid for a cartel to fix prices. If it did, colluders would set markups at the full cartel level; if it did not, they would not collude. This model, in which the probability does not depend on changes in λ, would make the empirical analysis somewhat more straightforward. However, such a formulation would be less descriptive of the actual situation confronting colluders. Formally, our assumptions do not preclude a corner solution at $\lambda = \lambda^c$. If the detection probability displays a limited responsiveness to markups, then eq. (2) may not be satisfied as an equality. However, it is unlikely, given the technology of collusion and the existence of punitive sanctions for price fixing, i.e., $t > 1$, for $(1 - dt)$ to be positive at $\lambda = \lambda^c$.

[11] See Sec. III; see also U.S. Office of the President 1969, and Elzinga and Breit 1976.

detection or the penalty for price fixing can be expected in most cases of interest to reduce the collusive markup.

II. Empirical Findings

Our theoretical model suggests that increases in enforcement levels or penalties for price fixing generally reduce collusive markups. Though straightforward to derive, this implication is hardly trivial to test. Unfortunately, we do not have a set of national, or even regional, industries having identical products, costs, and demand conditions but varying levels of antitrust enforcement. We must assess the impact of antitrust enforcement in a more problematic environment.

A. Choice of Industry

To test the implications of our deterrence model, we analyzed the market for white pan bread, a homogeneous commodity not only regionally produced and consumed but also well represented among DOJ price-fixing cases. During our sample period, bread cases were the most common among DOJ's food price-fixing cases.[12] In addition, the bread industry has well-recorded annual input and output prices compiled for selected cities by the Bureau of Labor Statistics (BLS) (see U.S. Department of Labor, Bureau of Labor Statistics 1964–76, 1965–76).[13] Hence, we constructed a sample that enabled us to use both cross-sectional and temporal variations in product prices, costs, and antitrust enforcement.

B. Estimating City-specific Markups

Developing an indicator for price markups in the bread industry involved two steps. First, we examined a standard recipe for a loaf of white bread.[14] We subtracted from each bread price observation the

[12] The FTC observed that the bread-baking and distribution industry has "the essential characteristic of a conspiracy-prone industry—relatively few sellers in individual markets" (U.S. Federal Trade Commission 1967, p. 135).

[13] The BLS publications contained comparable price data for white bread for 20 major cities. For 12 of these cities the data went back as far as 1955, although for the remaining eight they extended only to 1968. Input price data, though not so complete, proved adequate to construct a sample of 228 observations: 12 major cities for 1964–76, and eight additional major cities for 1968–76. Cities included throughout the entire period (1964–76) were Baltimore, Boston, Chicago, Cleveland, Detroit, Los Angeles, New York, Philadelphia, Pittsburgh, St. Louis, San Francisco, and Washington, D.C. For 1966–76, we were able to add Atlanta, Cincinnati, Dallas, Houston, Kansas City, Minneapolis, San Diego, and Seattle to the sample.

[14] The cost of ingredients (IC) for a 1-pound loaf of white bread is $\mathrm{IC} = .6350P_F + .0571P_S + .0026P_O + .035P_M$, where P_F is price/pound of flour, P_S is price/pound of

TABLE 1

EFFECT OF NONINGREDIENT INPUT COSTS ON ADJUSTED BREAD PRICE (PADJ)

Independent Variables	
PELEC	.256*
	(4.62)†
PGAS	2.49
	(3.15)
LABOR	1.07
	(7.14)
Constant	7.05
Observations (N)	228
R^2	.49
F-statistic (3,224)	72.93

* The estimated coefficient.
† The value of the estimated coefficient divided by its estimated standard error.

component attributable to the cost of the ingredients, IC, calculated from the recipe and from BLS ingredient prices. Formally, we defined the recipe-adjusted bread price, PADJ, as:

$$\mathrm{PADJ}_{it} \equiv p_{it} - \mathrm{IC}_{it}, \tag{5}$$

where i is the city and t the time index.[15] Differences in profits and other noningredient input costs cause PADJ to vary across cities and over time.

As a second step in developing a measure of markups, we estimated the amount of variation in PADJ attributable to variations in energy and labor costs. Table 1 contains the results of this analysis; PELEC and PGAS are measures of electricity and natural gas prices, and LABOR is a measure of wage rates for truck drivers. We assume this wage to be a proxy for general labor costs.[16] The results in table 1 suggest that variations in energy and labor input costs account for a reasonable amount of the variation in adjusted bread prices.[17]

sugar, P_O the price/pound of cooking oil, and P_M is the price/pound of dry milk. All prices are retail, except dry milk, which is wholesale. Cooking oil is used as a proxy for shortening (U.S. Executive Office of the President 1977, p. 7).

[15] The sources of price data were U.S. Department of Labor, Bureau of Labor Statistics (1965–76; 1969).

[16] Specifications using food and kindred workers' wages as well as alternative functional forms were also estimated. These different specifications yielded similar results (see Block et al. 1978).

[17] Of course, measures of goodness of fit for this regression substantially understate the importance of variations in input prices since the dependent variable, PADJ, is already net of the recipe or ingredient. Our entire procedure, taking account of both IC and the estimated cost function, accounts for approximately 82 percent of the variation in the price of bread (p). In general, prices for all significant inputs are included in cost functions. Our asymmetric treatment of input prices amounts to using

To construct a measure of the markup on bread, we then used the recipe costs and the results of this regression to estimate M, the markup indicator:

$$M \equiv [p - (\mathrm{NIC} + \mathrm{IC})]/(\mathrm{NIC} + \mathrm{IC}), \quad (6)$$

where all subscripts are suppressed and NIC refers to the fitted values generated by the equation in table 1.[18] Equation (6) estimates the markup by first subtracting from the market price the sum of the known ingredient costs (IC) and an estimate of other noningredient costs (NIC), and then dividing this difference by estimated unit costs. The higher the residual as a proportion of unit costs, the higher the markup.

C. *Estimating the Effect of DOJ Antitrust Enforcement on Collusive Markups*

Antitrust enforcement and penalties are only two of the many factors that actually determine markup levels.[19] Certainly, all factors that influence the price elasticity of demand also influence the markup level. Moreover, although we have assumed antitrust penalties to be the only cost of collusion, the resource cost of managing and policing a cartel also influences the optimal markup in the presence of antitrust enforcement. Assessing the effect of antitrust enforcement, therefore, requires a workable method of controlling for such outside influences. Influences generated by aspects of market structure that do not change rapidly—or at least that change significantly less rapidly than the antitrust variables—can be controlled by considering not the markup level but *changes* in the markup level.[20] We therefore used first differences in the markup level, or

$$\Delta M_{it} = M_{it} - M_{it-1}, \quad (7)$$

in testing the effectiveness of antitrust enforcement.[21]

---

a priori information—the recipe for bread—to supplant estimation of some coefficients. It may be argued that this procedure increases the efficiency of the estimation, but it precludes substitution among raw materials and other inputs.

[18] Because a simple regression generates NIC, approximately half the observations for NIC exceed PADJ and thus generate a negative M. For our purposes, these negative values pose no problems.

[19] We structured our test of antitrust enforcement around markups rather than the traditional—but indirect—test of overall profitability or rate of return. One reason for using the markup was the very directness of the test. Also, the availability of regional and city data enabled us to conduct more powerful tests of the effectiveness of antitrust enforcement than could have been performed with available data on rates of return.

[20] Controlling for these factors by actually assembling the relevant information on market structure for this industry at the city or SMSA level was not feasible.

[21] This procedure also facilitated our use below of a straightforward measure of regional antitrust enforcement.

1. Measuring Detection Probabilities, DOJ Enforcement Efforts, and Remedial Effects

We assume in our formal model of collusion that for any markup level the probability of detection is a function of DOJ enforcement efforts. Unfortunately, this straightforward theoretical proposition does not suggest a unique empirical counterpart for measuring either the probability of detection or DOJ's enforcement efforts. This formulation does suggest, however, that for any markup level the probability of detection is related to the capacity of the Antitrust Division to initiate cases. To the extent that the Division's litigation capacity relates to its expenditure level, the Division's annual budget should provide an indirect measure of this enforcement capability.[22]

A more direct measure of enforcement is simply DOJ price-fixing prosecutions. A price-fixing prosecution is rare enough for colluding firms in the affected industry to take special notice.[23] We have assumed, therefore, that each price-fixing case that the DOJ brings against a bread producer increases, for any markup level, the probability of prosecution perceived by other bread producers operating in the same DOJ region.[24]

We had sufficient price data to estimate changes in markups for 1965–76. We constructed a regional antitrust enforcement variable, DOJREG, for this period by setting the variable equal to one for each city within a region where the Antitrust Division filed an action that year—except for the city incurring the action—and by setting the variable equal to zero otherwise.[25] In other words, DOJREG is a shift

[22] A positive and significant relationship exists between the Antitrust Division's annual budget and the number of price-fixing cases brought by the Division. Using data on price-fixing cases during 1964–76 supplied by the Antitrust Division's Economic Policy Office, we estimated the relationship:

$$\text{DOJPF} = -11.39 + \underset{(2.42)}{.003\text{BUDGET}},$$

between the number of price-fixing cases brought annually by the Division (DOJPF) and its budget (BUDGET) measured in thousands of 1967 dollars (2.42 is the t-statistic).

[23] We can provide a formal rationale for this observation by assuming that colluders use Bayesian methods to estimate the probability that they will be apprehended in a particular period. In this formulation, whenever colluders are apprehended, colluders' estimate of the probability of apprehension increases, and that increase is dramatic if their a priori distribution is diffuse and has a small mean. Also, after an initial impact, the effect of the case on the apprehension probability estimated by the colluders deteriorates if no new cases are brought.

[24] The Antitrust Division operated seven regional offices during the years of our sample: New York, Philadelphia, Cleveland, Chicago, San Francisco, Los Angeles, and Atlanta. The Antitrust Division established a Dallas office in 1976 and realigned the areas covered by the regional offices. This study used the earlier seven regions for 1976 rather than realigning all regions for that single year. For coverage of regions, see U.S. Department of Justice, Antitrust Division (1973).

[25] This variable was constructed using data from the Commerce Clearing House (1955–75; 1966–76). A summary of those data appears in Block et al. (1978). Filing dates and other summary information were checked against a special listing of bread

variable designed to capture the changes in the perceived probabilities of cartel failure. Its form presumes that the colluders' estimate of the probability that DOJ will initiate a case for collusion increases after a DOJ price-fixing action within the same region.[26]

The variable DOJREM measures the effect of the antitrust action in the city where the Antitrust Division actually prosecuted a bread producer. In constructing DOJREM, we assumed that the impact of the antitrust action on the firms specifically prosecuted would differ significantly from the impact on firms in other cities in the same region. Basically we assumed that, for strategic reasons, the timing of the reduction in markups by prosecuted firms would not coincide with the reduction made by firms in nonaffected cities. To capture this effect, we set DOJREM equal to one in a city 1 year after the Antitrust Division had filed an action there.

2. Estimated Deterrent Effects

In table 2 we present our estimates of the effect of DOJ enforcement variables on markups in the bread industry. Again, the dependent variable, ΔM, is the annual change in the markup on white bread for the cities in our sample. The results in table 2 strongly suggest a deterrent effect of DOJ enforcement efforts.[27] First, the coefficient on

---

price-fixing cases prepared for us by the Economic Policy Office, Antitrust Division, U.S. Department of Justice, for the years 1963–76.

[26] An alternative hypothesis would be that, since DOJ has limited resources, a prosecution in a region signals that an additional prosecution in the same region is unlikely. Hence a firm actually would reduce its estimated probability of detection after a DOJ case in its own region. This hypothesis is equivalent to predicting that drivers on a highway accelerate when they see a car already being ticketed.

[27] Our formulation, which posits that increases in the markup increase the detection probability, generates the structural equations: $\Delta d = \xi + \beta\Delta\lambda + \tau\Delta\text{BUDGET} + \eta\text{DOJREG}$ and $\Delta\lambda = \psi + \alpha\Delta d + \theta\text{DOJREM}$. Therefore, the reduced forms are

$$\Delta d = \omega + \frac{\tau}{1 - \alpha\beta}\Delta\text{BUDGET} + \frac{\eta}{1 - \alpha\beta}\text{DOJREG} + \frac{\beta\theta}{1 - \alpha\beta}\text{DOJREM}$$

and

$$\Delta\lambda = \rho + \frac{\alpha\tau}{1 - \alpha\beta}\Delta\text{BUDGET} + \frac{\alpha\eta}{1 - \alpha\beta}\text{DOJREG} + \frac{\theta}{1 - \alpha\beta}\text{DOJREM}.$$

Our estimated reduced-form coefficients in table 2 are of the form $(\alpha\tau)/(1 - \alpha\beta)$, $(\alpha\eta)/(1 - \alpha\beta)$, and $(\theta)/(1 - \alpha\beta)$. We are mainly interested in deducing the sign of α given the sign of the reduced-form coefficients. The usual approach would be to estimate the other reduced-form equation in the system which takes Δd to be a function of the exogenous variable. Unfortunately this approach cannot be applied directly to this problem. We do not have a proxy for Δd or measures of the detection probability at a regional level. Even at a national level measuring d is problematic. However, given our formulation of the determinants of detection, if we assume that DOJ expenditures have some efficacy—i.e., $\tau > 0$—then the condition necessary to deduce the nonpositivity of α from the reduced-form coefficients is that $(1 - \alpha\beta) > 0$. This condition is required for Walrasian stability in the collusion market and in conjunction with our reduced-form

TABLE 2

Estimated Effects of Changes in DOJ Enforcement on Changes in Markups in the Bread Industry, 1965–76

Independent Variables				
ΔBUDGET	−.015*	−.015	−.024	−.020
	(−2.74)†	(−2.68)	(−4.06)	(−3.65)
DOJREG	−.025	−.026	−.025	−.027
	(−2.05)	(−2.21)	(−2.09)	(−2.26)
DOJREM		−.046	−.046	−.044
		(−2.32)	(−2.41)	(−2.32)
ΔFOODM			+.058	...
			(2.33)	
ΔGENM				−.010
				(−1.60)
Constant	.011	.013	.014	.017
R^2	.055	.082	.113	.101
F-statistic	5.93 (2,205)	6.04 (3,204)	6.47 (4,203)	5.68 (4,203)

Note.—Each regression is based on 208 observations.
* This coefficient is estimated per million dollars.
† The value of the estimated coefficient divided by its estimated standard error.

our general measure of changes in enforcement capacity—the change in the real value of the Antitrust Division's budget, ΔBUDGET—is negative and significant. In other words, an increase in the enforcement capacity of the Antitrust Division appears to reduce markups on white bread. Second, the coefficient on our direct measure of DOJ's enforcement activity (DOJREG) is negative and significant, suggesting that a price-fixing case against bakers in one city induces bakers in neighboring cities to reduce markups. This result comports both with our formal theoretical results and with the related conjecture in the Stigler report that "every victory" in seeking out price fixing "weakens the efficiency of undetected collusion."[28] Finally, the coefficient on DOJREM, the variable measuring the remedial effect of a price-fixing case, is negative and significant. Once discovered and prosecuted, colluders apparently "remedy" their price fixing by reducing their markups in the following year.[29]

results implies that there is a deterrent effect, i.e., $\alpha < 0$. In addition to this theoretical restriction on the sign of $(1 - \alpha\beta)$, we noted previously that there is a positive relationship between the total number of price-fixing cases brought by the Antitrust Division and the Division's budget level. If we can assume that other collusion-prone industries made decisions about collusion in the same way as bread producers, then this aggregate regression gives information about the second set of reduced-form coefficients. Again, if $\tau > 0$, the sign of the reduced-form coefficient on ΔBUDGET indicates that $(1 - \alpha\beta)$ is positive.

[28] See U.S. Office of the President, President Nixon's Task Force on Productivity and Competition 1969.

[29] It is only necessary to assume that DOJ capacity is productive ($\tau > 0$) to infer from our empirical results that $\eta > 0$, or that filing of a DOJ case in the region increases d.

The ΔFOODM and ΔGENM variables in table 2 control for general year-to-year variations in manufacturing markups that might not be adequately controlled by the first-difference procedure. The variable ΔFOODM is an annual series of the first differences in markups of food-and-kindred-product manufacturers; ΔGENM is an analogous series for all manufacturing firms.[30] The coefficient on ΔFOODM is of the expected sign, suggesting that, holding enforcement constant, markups in the bread industry move in the same direction as markups in other food-related industries. Although using the general manufacturing markup (ΔGENM) in the equation does not alter the estimates of the coefficients on the enforcement variables, the sign of the coefficient on ΔGENM is curious.

D. *Class Actions and the Effect of Antitrust Enforcement on Markups*

Historically, trial judges have punished price fixers leniently.[31] The cases collected for this study were no exception. All but two of the 17 bread price-fixing cases between 1957 and 1976 involved *nolo contendere* pleas, and in only one case did a defendant serve an actual prison sentence; moreover, total fines as a percentage of the pretax profits of the colluding firms averaged only 7 percent.[32] Neither imprisonment nor monetary penalties posed a credible threat to colluding firms. We hypothesized, therefore, that the deterrent effect of DOJ's enforcement efforts came not from the threat of publicly imposed fines or imprisonment, but from the increased likelihood of an award of private treble damages to bread consumers or distributors.[33]

[30] Both series were computed from data reported by the Federal Trade Commission in *Quarterly Financial Report for Manufacturing, Mining, and Trade Corporations,* in U.S. Council of Economic Advisers, *Economic Report of the President* (annual). Of course, both markup measures are influenced by antitrust enforcement efforts to the extent that collusion is important in these manufacturing industries.

[31] E.g., during 1966–76 over 85 percent of all price fixers who were convicted or who pleaded *nolo contendere* did not serve prison terms (see Block et al. 1978).

[32] For the 16 criminal cases involving either a *nolo* plea or a conviction the actual criminal fines for price fixing were:

	Minimum	Maximum	Average
Individuals ($)	1,031	9,966	4,025
Firms ($)	2,930	50,638	20,690
Fines, as % of defendants' annual sales	.09	1.0	.31

All fines are in 1976 dollars. Data on the ratio of fines to sales were derived from court documents relating to 13 cases (76 firms) during 1957–75. The overall ratio of net profits to sales was obtained from a study by the U.S. Executive Office of the President, Council on Wage and Price Stability (1977).

[33] Civil actions alleging horizontal price fixing are possible but, certainly in this industry, uncommon without a preceding criminal case.

Since bread price fixing generally causes a small injury to many individual distributors and consumers, private damage recovery is usually feasible only through a class-action suit. Class actions enable plaintiffs who are numerous, and whose independent damage claims are too insignificant to justify litigation, to maintain a single action for their aggregate damages.[34] District court documents revealed that settlements in class actions for price fixing in the bread industry were almost 10 times greater than government-imposed fines.[35]

To test the hypothesis that private class actions actually provided the effective penalty in price-fixing cases, we partitioned our sample into the periods before and after class actions became a credible threat in the bread industry. Since only one class action in our sample did not follow a DOJ case, we assumed in partitioning the sample that

[34] These actions, brought under Rule 23 of the Federal Rules of Civil Procedure, have become far more frequent since the Supreme Court amended that rule in 1966. The amended Rule 23 contributed to this greater frequency by changing the procedure for becoming a member of the class. Originally, Rule 23 required persons to "opt into" the class before they could benefit from the adjudication of the class's cause of action; the amended rule instead presumed persons to be class members unless they affirmatively "opted out" of the class. Although surprisingly few historical data exist on class actions, the data that do exist—particularly the information on docket entries for the Southern District of New York collected by the American College of Trial Lawyers' Special Committee on Rule 23—suggest that the 1966 amendments to Rule 23 made the class action a much more attractive legal device (American College of Trial Lawyers 1972). A recent alternative to the antitrust consumer class action is the *parens patriae* device by which a state attorney general may sue on behalf of the consumers in his state. For the purposes of deterrence, *parens patriae* actions are virtually identical to class actions (see Block et al. 1978).

[35] A search of Commerce Clearing House (1955–75; 1966–76), McLaughlin (1976), Newberg (1977), and the LEXIS computer file (1978) revealed seven major class actions for price fixing in the bread industry since 1966. According to district court documents, class-action settlements or damages during 1971–76 obtained for price fixing in the bread industry were:

	Minimum	Maximum	Average
Cases ($)	1,197,810	6,100,000	1,998,646
Firms ($)	39,562	1,220,000	293,919
Damages as % of defendants' annual sales	.41	19.68	2.87

All fines are in 1976 dollars. These figures are based on data for five of the seven recorded cases. One of the seven original class actions had no settlement or award since the district court refused to certify the class. No data at all were available on another case. All five cases were settled rather than litigated to judgment. Details of these class actions appear in Block et al. (1978, appendix table VI). Whereas average damages as a percentage of defendants' sales were only 10 times average fines as a percentage of defendants' sales, average damages per firm were almost 15 times average fines per firm. This occurred because the average sales per defendant differed significantly between the two samples: $5,795,271 for the fine calculations and $10,254,205 in the class-action calculations.

TABLE 3

ESTIMATED EFFECT OF DOJ ENFORCEMENT AND CLASS ACTIONS ON MARKUPS IN THE BREAD INDUSTRY, 1965–76

Independent Variables		
ΔBUDGET11	−.018* (−.622)†	. . .
ΔBUDGET12	−.014 (−2.68)	. . .
ΔBUDGET21	. . .	−.002 (−.158)
ΔBUDGET22	. . .	−.019 (−3.53)
DOJREG11	+.004 (.160)	. . .
DOJREG12	−.037 (−2.70)	. . .
DOJREG21	. . .	−.019 (−1.03)
DOJREG22	. . .	−.029 (−1.87)
DOJREM	−.046 (−2.43)	−.042 (−2.19)
Constant	.013	.013
Observations (N)	208	208
R^2	.0926	.1024
F-statistic	4.13 (5,202)	4.60 (5,202)

* This coefficient is estimated per million dollars.
† The value of the estimated coefficient divided by its estimated standard error.

class actions affected primarily the penalty cost of detection, not the probability of detection.[36] We considered as partition dates the year that a district court first certified a class in a bread case (1970) and the year that the Administrative Office of the Courts began reporting class-action activity (1972). Before 1970 class actions probably were not a credible threat in the bread industry; by 1972 no doubt remained that class actions had become prevalent and important. To accomplish the actual partitioning of the sample, we split the enforcement variables ΔBUDGET and DOJREG around 1970 and 1972. In table 3, ΔBUDGET11 and DOJREG11 are ΔBUDGET and DOJREG for 1965–69, while ΔBUDGET12 and DOJREG12 are the variables for 1970–76. The comparable variables partitioned around

[36] To determine when, after the 1966 amendments to the class-action provisions in the Federal Rules of Civil Procedure, the class action became a credible threat in the bread industry, we searched for both recorded and unrecorded class actions involving horizontal price fixing of bread products. None of the seven class actions we found was filed before 1968, actually certified before 1970, or settled before 1971.

1972 instead of 1970 are ΔBUDGET21, ΔBUDGET22, DOJREG21, DOJREG22.[37]

The estimates in table 3 are consistent with our hypothesis that class actions represent the effective penalty in price-fixing cases. This result does not seem to depend upon whether we partition the enforcement variables around the date of the first class certification or around the date of the first reporting of class-action activity by the Administrative Office of the Courts. For either partition, only in the latter period, when class actions represented a credible threat, did a significant deterrent effect result from either an increase in the Antitrust Division's resources or from the actual prosection of a horizontal price-fixing conspiracy.[38]

The pattern in the bread industry is clear. A successful federal prosecution signals to consumers that a treble damage suit has become feasible. Private plaintiffs subsequently provide the effective penalty in the form of class actions for treble damages.[39] For price fixing in this conspiracy-prone industry—and, we suspect, for price fixing in general—deterrence has been a product of both public and private enforcement efforts.[40]

[37] Formally, the definitions of the variables are: ΔBUDGET21 = ΔBUDGET, 1966–71, zero otherwise; ΔBUDGET22 = ΔBUDGET, 1972–76, zero otherwise; DOJREG21 = DOJREG, 1966–71, zero otherwise; DOJREG22 = DOJREG, 1972–76, zero otherwise.

[38] While the estimates in table 3 do not include the markup controls, ΔFOODM and ΔGENM, regressions with these controls yielded similar results. Several factors possibly confound this analysis of the deterrent effect of class actions. First, the average criminal fine imposed by district courts might have increased over the period, or might simply have been larger when class actions became a relevant concern. To check for this possibly confounding influence, we estimated time trends for several measures of criminal fines. Overall, the evidence suggests that an increase in criminal penalties did not confound our results on the deterrent effect of class-action suits. Second, the early 1970s included a period of price controls, and DOJREG12 and/or ΔBUDGET12 possibly proxied for the effect of these controls. We tested whether the enforcement variables performed this role by including directly in the markup regression a dummy variable for price controls. The results of this procedure suggest that, although price controls had a significant depressing effect on markups, the enforcement variables were not merely a proxy for price controls. Finally, we controlled for pure time trends associated with changes in demand or market structure within each city. This was accomplished by regressing the dependent variable (ΔM) against city dummies along with the deterrent variable. The effect of the enforcement variables remained significant and unchanged and nearly all of the city dummies were statistically insignificant (see Block et al. 1978).

[39] A *nolo* plea apparently is a sufficient signal. Although guilty verdicts are prima facie evidence in a treble damage suit and should induce private enforcement, they are rare in price-fixing cases. In our sample of 17 price-fixing cases, one case ended in an acquittal and another in a conviction; the other 15 involved *nolo* pleas. Yet even a government case that ends in a *nolo* plea signals to potential private plaintiffs that they probably would prevail in a damage suit against the alleged price-fixing conspiracy. In fact, all the reported class actions in our sample that followed a government action and received class certification eventually obtained a settlement award.

[40] The amendment of Rule 23 in 1966, of course, was the event which facilitated class actions, thereby increasing the power of the federal government to deter price fixing.

III. Conclusion

In this paper we formulated and tested a simple model of collusive pricing in the presence of antitrust enforcement. We showed that if a cartel's probability of detection increases with its markup, then the cartel's optimal price is neither the competitive price nor, in most cases, the price that a cartel would charge in the absence of antitrust enforcement, but rather an intermediate price that depends on the levels of antitrust enforcement efforts and penalties.

Our empirical results revealed that increasing DOJ's enforcement capacity or filing a DOJ price-fixing complaint had the deterrent effect of reducing markups in the bread industry. We noted that government-imposed price-fixing penalties were trivial and found support for the proposition that the effective deterrent to price fixing was the credible threat of large damage awards to private class actions that followed DOJ's case against the same conspiracy. Consequently, only after class actions became a credible private remedy did the Antitrust Division's enforcement capacity or its filing of a bread price-fixing case deter collusion in the conspiracy-prone bread industry.

References

American College of Trial Lawyers. *Report and Recommendations of the Special Committee on Rule 23 of the Federal Rules of Civil Procedure.* Los Angeles: American Coll. Trial Lawyers, 1972.

Block, Michael K.; Nold, Frederick C.; and Sidak, J. Gregory. "The Deterrent Effect of Antitrust Enforcement: A Theoretical and Empirical Analysis." Technical Report no. ISDDE-1-78, Center for Econometric Studies of the Justice System, Hoover Institution, Stanford Univ., December 1978.

Burnham, David. "Tougher Penalties in Antitrust Cases." *New York Times* (March 16, 1978).

Commerce Clearing House. *Trade Regulation Reporter.* New York: Commerce Clearing House, various issues, 1955–75.

———. *Trade Cases.* New York: Commerce Clearing House, various issues, 1966–76.

Elzinga, Kenneth G., and Breit, William. *The Antitrust Penalties: A Study in Law and Economics.* New Haven, Conn.: Yale Univ. Press, 1976.

Hay, George A., and Kelley, Daniel. "An Empirical Survey of Price Fixing Conspiracies." *J. Law and Econ.* 17 (April 1974): 13–38.

McLaughlin, Joseph T., ed. *Federal Class Action Digest 1976.* New York: Practising Law Inst., 1976.

Landmark Supreme Court decisions regarding class-action suits such as Eisen v. Carlisle and Jacquelin (417 U.S. 156 [1974]) and legislation such as the Hart-Scott-Rodino Antitrust Improvement Acts are likely to have similarly far-reaching effects. Some very preliminary attempts to assess the impact of these events on collusion appear in Block et al. (1978).

Newberg, Herbert B. *Newberg on Class Actions: A Manual for Group Litigation at the Federal and State Levels.* New York: McGraw-Hill, 1977.

Posner, Richard A. "A Statistical Study of Antitrust Enforcement." *J. Law and Econ.* 13 (October 1970): 365–419.

Stigler, George J. *The Organization of Industry.* Homewood, Ill.: Irwin, 1968.

U.S. Council of Economic Advisers. *Economic Report of the President.* Washington: Government Printing Office, annual.

U.S. Department of Justice, Antitrust Division. "Memorandum to All Personnel. Re: Organization and Operation of the Antitrust Division." Directive no. 10-73, September 30, 1973.

U.S. Department of Labor, Bureau of Labor Statistics. *Wholesale Prices and Price Indexes.* Washington: Bur. Labor Statis., various issues, 1964–76.

———. *Estimated Retail Food Prices by City.* Washington: Bur. Labor Statis., various issues, 1965–76.

———. *Retail Prices of Food, 1964–68.* Bulletin no. 1632. Washington: Bur. Labor Statis., 1969.

U.S. Executive Office of the President, Council on Wage and Price Stability. *A Study of Bread Prices.* Washington: Executive Office of the President, April 1977.

U.S. Federal Trade Commission. *Economic Report on the Baking Industry.* Washington: Federal Trade Commission, 1967.

U.S Office of the President, President Nixon's Task Force on Productivity and Competition. *Recommended Changes in Antitrust Policies.* Washington: Government Printing Office, 1969.

The Market Evaluation of Human Capital: The Case of Indentured Servitude

David W. Galenson

University of Chicago and California Institute of Technology

This paper examines the market for human capital created by the institution of indentured servitude in colonial America. The indenture system allowed English emigrants to obtain passage to the colonies by selling claims on their future labor. With the size of the debt approximately equal for all emigrants, the length of the term for which a servant was bound is predicted to have varied inversely with expected productivity in the colonies. Analysis of two collections of contracts made in the seventeenth and eighteenth centuries supports the prediction. Age, skill, and literacy were negatively related to length of indenture. Women received shorter terms than men at young ages, while servants bound for the West Indies and those bound in periods of high colonial demand for labor also received reductions.

I. Introduction

During the colonial period of American history, two institutions existed which provided for the explicit valuation of stocks of human capital in the market. One of these was slavery, under which blacks and their progeny were held in service for life. The other was indentured servitude, under which whites were bound to service for limited periods of time.

I am grateful to Stanley Engerman for discussions of many of the issues treated in this paper and comments on an earlier draft. I would also like to thank Andrew Abel, Gary Becker, Lance Davis, Robert Fogel, Richard Freeman, Russell Menard, Frederic Mishkin, Melvin Reder, Sherwin Rosen, T. W. Schultz, George Stigler, and participants in seminars at Columbia University, the University of Chicago, the University of Illinois, the Newberry Library, the University of Southampton, and the 1980 Cliometrics Conference for their suggestions and comments.

[*Journal of Political Economy*, 1981, vol. 89, no. 3]

As a result of both its direct importance for the history of the nineteenth century and its indirect importance for the history of the twentieth, slavery has long been one of the central concerns of American social and economic historians. However, in the labor markets of many regions, for substantial periods during the first half of the colonial era, indentured servitude was of greater quantitative importance than slavery.[1] In the eighteenth century the quantitative significance of indentured servitude declined in most of these areas, as slaves were substituted for servants in the sugar fields of the West Indies, in the tobacco fields of the Chesapeake colonies, and in the rice fields of South Carolina.[2] Yet indentured servitude nonetheless continued to perform an important role in the colonial labor market by providing skilled craftsmen and managers to the large plantations of the West Indies and the southern mainland colonies.[3]

The essential difference between servitude and slavery was that it was the labor of the servant, rather than the person, which was bought and sold. Because less attention has been devoted to the study of indentured servitude, we are less familiar with the history of the institution. An economic investigation of its functioning, based on the analysis of quantitative data generated by the operation of the market for servants, can therefore serve a dual purpose: It can provide evidence of how capital values were established when claims on the labor of humans for long periods were traded in the market, while at the same time increasing our understanding of the economic basis of white servitude in colonial America.[4]

II. The Indenture Bargain

Indentured servitude was a credit system under which human labor was leased. It functioned through two markets linked by a recruiting agent. In England, in the first market, a prospective servant signed a

[1] An example is Maryland, where a study of probate inventories has indicated that as late as 1674–79 the ratio of servants to slaves held in estates was 3.88 (Menard 1977*b*, p. 360). It should be pointed out that in some areas neither type of bound labor was quantitatively important. Slaves never accounted for more than 3 percent of New England's total population in the colonial period, and the share of servants was probably of a similar magnitude (Greene 1942, chap. 3; Abbot Emerson Smith 1947, pp. 28–29). A generalization that held for a number of major colonial regions is that the share of bound workers in a region's labor force tended to rise as one traveled south from New England, with progressively increasing shares in the middle colonies, the Chesapeake, South Carolina, and the West Indies. For evidence on population composition, see Greene and Harrington (1932), Sutherland (1936), and Wells (1975).

[2] For accounts of this process in the West Indies, see Dunn (1973); for the Chesapeake, see Menard (1977*b*); for South Carolina, see Wood (1975).

[3] On this change, see Gray (1958, 1:350) and Pares (1960, p. 19); for additional evidence see Galenson (1979*a*, chap. 10).

[4] The present paper extends the analysis and empirical findings of Galenson (1977*b*).

contract, or "indenture," with a merchant, promising to serve the latter or his assignees in a particular colony for a given period under stated conditions. The servant was then transported to the specified colonial destination, where the merchant or his representative sold his contract to a colonial planter or farmer in the second market. In return for the commitment of his labor, the servant received passage to the designated colony, maintenance during the term of the contract, and certain freedom dues at its conclusion. Once signed, the indenture was negotiable property, and at any time before its conclusion the servant could be sold to a new master for the balance of his term. When the contract expired, the servant became free. The conditions of servitude were regulated by colonial statutes as well as by agreements written into the contracts. The terms of the contract were binding upon both master and servant.

The indenture system normally operated within a context of competitive markets both in England and in the colonies. Servants were one important available backhaul cargo for English ships engaged in the trade for colonial sugar, tobacco, rice, and other agricultural staples.[5] Contemporary fare quotations indicate that the charges for passage from England to America were uniform at a given time for all individuals and did not vary by specific colonial destination.[6]

The planter's demand for indentured servants was based on his calculation of the discounted value of their net future earnings, after deducting the expected costs of the servant to him. The present value of the servant to the planter therefore depended upon the expected value of the servant's output in each year of the contract; the expected cost of maintenance, supervision, and training for the servant during each year of the term; the discount rate; and the value of the freedom dues to be paid to the servant. These, or analogous variables, are the same considerations which enter into the derivation of a free worker's net age-wealth profile (Becker 1975, p. 223). Yet one critical difference is that, whereas for free workers evidence on flows is used to calculate the values of capital stocks, in the case of servants these calculations were performed by planters, who based their demand for stocks of bound labor for fixed terms on their calculations involving

[5] See Bruce 1907, 1:622; Smith 1947, p. 39; and Middleton 1953, pp. 145–56.

[6] While £6 was the fare cited in the early colonial period, after the middle of the seventeenth century £5 was the fare normally quoted for passage to all colonies. For references to quotations of passage charges from England to a number of colonies, see, e.g., John Smith (1624, pt. 1, p. 162); Purchas (1625, pt. 4, p. 1791); Bullock (1649, p. 47); Taunton (January 4, 1670); Wilson (1682, p. 19); Littleton (1689, p. 17); Jeaffreson (1878, pt. 2, p. 102); Kingsbury (1906, pt. 1, pp. 277–78); Abbot Emerson Smith (1947, p. 35); and Alexander (1972, p. 45). A qualification to the statement in the text is possible seasonal variation in the cost of delivering servants to some colonies, discussed below.

the relevant flows. Therefore, like the slave market, the market for indentured labor produced capital values, and the flows underlying these must be inferred.

The institutional arrangements which provided for the ownership of human capital for discrete periods produced potential differences between the patterns of human-capital values under indentured servitude and those implicit in the operation of free labor markets. An advantage of the indenture system for the planter, relative to the hiring of free workers, was the greater control it gave him over the servant's time and effort once the bargain was made. Because the servant would have been compensated for his loss of freedom, in the absence of uncertainty, the price of indentured human capital would have been higher than that of free workers.[7] However, in practice, uncertainty makes the direction of the inequality between these prices unpredictable without additional information, for it cannot be determined a priori whether the insurance value of the contracts was normally greater to the planter or the servant.[8]

All servants who migrated to America incurred debts of similar value. As noted above, passage charges were uniform for all servants, and maintenance costs and freedom dues varied little across individuals. As a result, every servant contract was a promise to repay approximately the same sum of money. Therefore, the present discounted value of every servant's contract should have been approximately the same at the time of binding. Since the productivity of servants varied, the conditions of their indentures had to vary. The higher ther servant's marginal-value product above his maintenance, the faster he could repay the loan made to him, and the shorter the term of the contract. The length of indenture across servants should, therefore, have been negatively correlated with individual productivity or, equivalently, with the market valuation of the current flow of income generated by the individual's stock of human capital.

[7] This assumes a solution to the problem of monitoring the servant's work in which the productivity gains from the master's control over the servant's time and effort were not offset by shirking.

[8] In a world in which futures contracts for free wage labor could be made with certain fulfillment, the present value of an indenture for a given number of years and of a series of contracts for hires for the same years would differ only by a premium which would reimburse the servant for the loss of freedom resulting from his residence in the master's household and due to other legal provisions governing servitude. However, in the absence of these guaranteed futures contracts, in some cases the master might also have been willing to pay more for an indenture because of the assurance it gave him of labor supply in peak seasons or future years; similarly, workers might sometimes have been willing to accept lower implicit wages in return for the guarantee of employment the indenture represented. The relative cost of this insurance to master and servant could vary, and as a result the relation between the implicit wage paid to servants and the hire rate for free workers is indeterminate.

Strictly, if costs had been precisely the same for all servants, all should have been bound on conditions that would yield the same expected price on arrival in the colonies. This, of course, does not imply that no variation should have occurred in the realized auction prices of servants in the colonies, for random disturbances between the time of binding and auction—such as illnesses of servants during the voyage or changes in the price of colonial outputs—could have produced differences between actual and expected prices. What the analysis does imply, however, is that if the market for contracts of servitude was efficient, variation in servants' auction prices should have been uncorrelated with all the characteristics of the servants which were known at the time of making the indenture bargain.[9]

Very little evidence of the prices of newly arrived servants at colonial auctions has survived: From the entire colonial period the only known records of auctions of English servants come from the accounts of two ships, the *Tristam and Jeane,* which sold 68 servants in Virginia in late 1636 or 1637, and the *Abraham*, which sold 56 servants in Barbados in January 1637 (London, Public Record Office, 1636, 1637). The listings of the amounts paid for the servants show that the median and modal prices of both men and women were identical—500 pounds of tobacco—in both auctions. Thus, although as will be seen both sex and colonial destination had a significant effect on the length of servants' indentures, the available evidence, though limited, suggests that neither may have affected the initial colonial auction prices for the servants. This is consistent with the hypothesis that the variable dimensions of the contract were adjusted for the servants' characteristics so as to make the expected auction prices of all servants the same.

The relation between auction prices of servants and the simple cost of passage is also of interest. The modal price of 500 pounds of tobacco observed in the auctions was 11 percent greater than the fare that two free passengers each paid for passage to Virginia on the *Tristam and Jeane* on the same voyage as the servants and 4 percent above the fare quoted elsewhere for passage from England to Maryland in 1638.[10]

[9] This statement neglects one potential element of cost. The marginal cost of delivery, and therefore auction prices, would have included any costs the merchant incurred in recruiting servants in England. These could clearly vary across individual servants, producing differences in expected auction prices. These costs could have been correlated with individual productivity, as in some instances skilled servants were given lump-sum payments at the time of binding. For some evidence of this practice, see the analysis of the Middlesex sample in the Appendix.

[10] The two free passengers each paid 450 pounds of tobacco (London, Public Record Office 1637). For the 1637 Chesapeake farm price of tobacco, see Menard (1975, p. 475). The fare quotation, of £6 sterling per person, is in Maryland Historical Society (1889, p. 206).

That in these cases the typical auction prices tended to be higher than the fare may have been due to the cost of recruiting servants, or to the existence of a premium received by merchants for bearing the risk of servant mortality on the ocean voyage.

Related evidence on the nature of the labor-market equilibrium in the colonies which induced flows of labor from England is available from surviving valuations of indentured servants recorded in Maryland for probate courts. The mean price of 28 male servants with 4 remaining years of servitude recorded on Maryland's lower western shore during 1704–57 was £8.95, with a median of £9 and a mode of £10; while the mean price of 19 females with 4 years remaining was £7.75, with a median of £8 and a mode of £10.[11] The typical marginal cost to an English merchant of delivering servants to the colonies cannot be estimated precisely but probably fell within a range bounded at the lower end by £5, the usual cost of passage, and at the upper end by the £10 estimated by Abbot Emerson Smith as the maximum expense of delivery.[12] The evidence of the probate valuations—although again limited in quantity—suggests that the central tendency of the colonial price of servants did fall within this range and, therefore, offers additional support for the hypothesis that the price of a servant's indenture in the colonies was equal to the marginal cost of delivering labor there from England.

A number of other general considerations relating to the adjustment of the bargain deserve mention. One is the role of mortality. The smaller the probability of a servant's serving a given year of his term, the lower his expected net earnings and, ceteris paribus, the less favorable the terms of the contract he would be able to sign. This is true whether the mortality in question is that during the Atlantic crossing, when the merchant bore the risk, or that after arrival in the colonies, when the planter had purchased the contract and assumed the risk.[13]

A second factor with a similar effect was the possibility of a servant failing to serve out his term for a reason other than death, principally

[11] I am grateful to the St. Mary's City Commission, Annapolis, Maryland, for transcriptions of the probate price quotations, from Maryland Hall of Records. Most of the quotations are from the early decades of the period; on devaluation of Maryland currency and its reflection in probate valuations, see Main (1972, pp. 14–18) and McCusker (1978, pp. 189–204).

[12] Smith 1947, p. 37. A precise analysis would include a positive premium in the probate valuations, for these servants had normally been "seasoned," i.e., had spent a year in the new colonial disease environment, and consequently had a longer life expectancy than the new arrivals, ceteris paribus.

[13] Mortality rates differed among colonies, and the effects of this on the length of indenture will be discussed below. The assumption here will be that mortality rates among servants both during passage and in the colonies were not systematically related to individual productivity.

running away. All the colonies enacted legislation intended to discourage servants from running away; of these some were preventive measures and others punitive. While some colonies provided for corporal—and even capital—punishment for runaways, the most common penalty was extension of the servant's contract by some multiple of the time he was absent.[14] One of the provisions of the contract, the servant's freedom dues, constituted a nonvested pension and, therefore, also acted to discourage servants from running away. While the form and worth of the dues varied across colonies and over time, they were often of substantial value and could have constituted a significant deterrent to servants who considered escaping from their masters.[15]

Servants were not allowed to marry during their terms. Since, by English practice, the expense of raising the illegitimate children of servants fell on the county, colonial legislation provided that the father should be discovered by oath of the mother and that he should reimburse the county for the expense of raising the child until it could be bound out to work. Since servants could not normally pay this sum, either the master would pay it and the servant's term would be extended, or the servant would be bound over to the county for an additional term after the conclusion of his normal term, to be sold for the necessary amount. The mother's term was also extended to reimburse the master for her lost working time.

III. Empirical Results

The principal sources of quantitative evidence bearing on this market for human capital are servant contracts recorded and held in English

[14] For references to laws relating to runaways, see Semmes (1938, pp. 116–18) and Smith (1947, pp. 264–70). The strong terms of this colonial legislation were an important element in the indenture system's success, and some authors have argued that changes in the legal provisions for the enforcement of contracts were central to its decline in the early nineteenth century; see, e.g., Geiser 1901, p. 42. Similarly, the lack of success of the attempt to revive a contract labor system during 1864–85 has been attributed in part to the high cost of enforcing contracts through civil action; see Erickson 1957, pp. 46–48.

[15] On freedom dues, see Smith (1947, pp. 238–41) and Heavner (1976, pp. 50–51). For a discussion of an analogous provision, nonvested pensions as a firm's insurance against quits, see Becker (1975, p. 34). The nature of freedom dues has some implications for other dimensions of the servant contract. The dues were specified by colonial law and were equal for all servants in a given colony. Therefore, the discounted cost of the dues to the planter at the time he purchased a contract varied inversely with the length of the contract. One effect of the fixed nominal value of the freedom dues may, therefore, have been to reduce the amount of variation in the length and other dimensions of the contracts, since the variations in the latter were intended to equalize the net present values of all contracts. This point should not obscure another basic effect of the existence of freedom dues, for, ceteris paribus, they raised the cost of servants and therefore tended to lengthen the term of indenture.

courts. These were made both to protect servants from kidnapping and to protect merchants from false charges of kidnapping. The two largest known surviving collections will be analyzed here; the earlier set, recorded in Middlesex during 1683–84, covers a total of 812 individuals, while the later, recorded in London during 1718–59, covers 3,187 servants.

The principal variable analyzed here will be the length of the indenture. Furthermore, this analysis will be done only for the minors in both samples; these comprise 22 percent of the servants with known ages in the earlier and 67 percent in the later sample. The selection of the variable to be analyzed and of the minors follows from consideration of the evidence of the contracts. It is clear that 4 years was the normal duration of an adult's indenture in both samples, yet for a combination of legal and clerical reasons the full set of conditions of adults' servitude do not appear to have been recorded in either set of indentures. For minors, the full conditions do appear to have been recorded, and the duration of the contract appears to have been the chief variable dimension of the contracts. Both variations in freedom dues and restrictions on the servants' occupations in the colonies were rare. Cash payments made to adults were not generally recorded in the later sample, but the contracts of the minors, on which they do appear to be recorded conscientiously, show that fewer than 6 percent received cash payments. Some cash payments are recorded for both minors and adults in the earlier sample, and although it is uncertain whether payments were recorded in all cases in which they were made, less than 7 percent of the contracts contain such entries. A notable feature of the payments recorded in both samples is that virtually all—95 percent in the earlier and 97 percent in the later sample—were made to servants bound for 4 years. Insofar as payments were made (or promised) and recorded, analysis of the characteristics of servants who received them tends to reinforce the results obtained from the analysis of the length of indenture.[16] The evidence of both samples, therefore, indicates that for minors the greatest variation in the conditions of indenture occurred in the duration of the term of servitude, while for adults this was not the case, as 4 years was both the standard term for adults and normally the minimum term assigned.[17] For adults, cash payments were appar-

[16] See the Appendix.

[17] Terms of less than 4 years do occur, but they appear to have been rare after the midseventeenth century; thus, they account for only 0.5 percent of all indentures of known length in the Middlesex registrations of 1683–84 and 1.2 percent of those in the London registrations of 1718–59. The reasons for this are not known. Four years may have been the term required at most times and places for the average adult to repay the cost of passage out of his net earnings, but it is unclear why shorter terms were not more often given to highly skilled servants like the accountant James Corss, whom

ently substituted for reductions in the length of servitude below 4 years.

Earlier analysis suggested that a servant's term of indenture would be inversely related to the market valuation of his stock of human capital. This index of the servant's human capital can be related to a number of observable characteristics potentially relevant to the determination of the present value of that stock. When this is done by multiple-regression analysis, the estimating equation differs from the common hedonic method only in the use of an index for price. The estimated coefficients of the independent variables age, sex, literacy, and occupation can be interpreted as the marginal prices paid for servants' characteristics in units of the index, while those of destinations represent compensating differentials among regions.[18]

Table 1 shows a number of the basic relationships underlying the market valuation of the servants. In both samples, the length of indenture was negatively related to both age and skill: With other things equal, servants with skilled occupations and those able to sign received shorter terms. Women received shorter terms than men, ceteris paribus, and servants bound for the West Indies received shorter terms than those bound for the North American mainland.

A comparison of the estimated coefficients of the sex variable across samples indicates that, on average, women received considerably larger reductions in their terms in the 1680s than in the eighteenth century. This decline in the premium for females is not surprising in view of the generally declining colonial sex ratios during this period, for while women were preferred for some kinds of household work and some types of farming, their increasing relative availability in most colonies would be expected to lower the size of their wage differentials.[19]

Table 2 provides a more detailed analysis of the length of indenture for the later sample, allowing separate age profiles of length of indenture by sex and skill. It reveals that there was a tendency for women to receive indentures from 5 to 15 percent shorter than those of men through the age of 17, while for servants aged 18–20 there was no

---

Walter Tullideph sent to the manager of his plantation in Antigua in 1759 with a note stating that he "hath bound himself to serve me four years agreeable to the Laws of Antigua, but as he is 22 years of Age, he thought it hard to serve so long and for that reason, I have given him a Certificate that he is to be absolved from the last year's Service" (Tullideph 1759, vol. 3). That planters preferred to substitute salaries for reductions of the term below 4 years suggests the possibility that fixed costs of hiring and/or a desire to capture the returns from a servant's general training in the colony may have been important considerations.

[18] On the interpretation of coefficients in hedonic price indexes, see Rosen (1974, pp. 34–35).

[19] The difference between samples in the sex coefficients in table 1 is significant at the .01 level. On declining colonial sex ratios, see Wells (1975, pp. 156, 219, 244).

TABLE 1

ESTIMATED REGRESSION COEFFICIENTS, MIDDLESEX AND LONDON SAMPLES

	Middlesex, 1683–84		London, 1718–59	
Independent Variable	Estimated Coefficient	Standard Error	Estimated Coefficient	Standard Error
Age (years):[a]				
Total sample:				
Less than 15	2.655	.385	2.749	.134
15	2.201	.400	2.147	.080
16	1.457	.304	1.304	.068
17	.893	.367	.728	.062
18	.174	.270	.331	.055
19	.738	.306	.169	.050
Sex[b]	−1.484	.207	−.195	.073
Literacy[c]	−.575	.217	−.082	.037
Date[d]	...	...	−.0060	.0023
Trade[e]	−.727	.445	...	...
Farmer[f]	...	...	−.313	.074
Laborer	...	...	−.146	.079
Services[g]	...	...	−.348	.066
Metal and construction[h]	...	...	−.320	.067
Clothing and textiles[i]	...	...	−.313	.060
Antigua[j]	−.227	.812	−.403	.110
Barbados	−.553	.274	−.176	.154
Jamaica	−.398	.462	−.233	.060
Other West Indies[k]	−.401	1.094	−.479	.088
Maryland	.203	.209	.306	.059
Virginia	...	...	.127	.073
Other mainland[l]	−.389	.673	.050	.116
Constant	5.227	...	4.665	...
R^2	.555	...	.539	...
F	12.87	...	112.82	...
n	171	...	2,049	...

SOURCE.—Data used are all from records of minors (age less than 21). Middlesex, 1683–84: London, Greater London Record Office; Nicholson (1965); Wareing (1976). London, 1718–59: London, Corporation of London Records Office; Kaminkow and Kaminkow (1964); Galenson (1977*a*).

NOTE.—Dependent variable = number of years indentured; method of estimation used is ordinary least squares in tables 1 and 2.

[a] For age variable, indicated age = 1; zero class = age 20.

[b] Male = 0, female = 1.

[c] Marked = 0, signed = 1.

[d] Date entered as final two digits of year of registration.

[e] Trade = 0 for laborers and no recorded occupations; trade = 1 for all other men's occupations.

[f] For all occupational variables, indicated occupation(s) = 1; zero class = no recorded occupation. "Farmer" includes husbandman, plowman, etc.

[g] Includes accountant, barber, surgeon, etc.

[h] Includes blacksmith, carpenter, cooper, mason, etc.

[i] Includes clothier, tailor, weaver, etc.

[j] For all destination variables, for Middlesex sample, zero class = Virginia; for London sample, zero class = Pennsylvania.

[k] Includes Nevis, St. Christopher, etc.

[l] Includes Carolina, New York, etc.

TABLE 2

ESTIMATED REGRESSION COEFFICIENTS, LONDON SAMPLE, 1718–59

Independent Variable	Estimated Coefficient	Standard Error
Age (years):		
Total sample:		
Less than 15	2.976	.144
15	2.378	.092
16	1.542	.084
17	.959	.084
18	.473	.079
19	.260	.083
Women, age:[a]		
Less than 15	−1.034	.460
15	−.472	.390
16	−.969	.228
17	−.302	.161
18	.041	.159
19	.090	.135
20	.198	.156
Trade, age:[b]		
15	−1.496	.392
16	−.884	.206
17	−.502	.117
18	−.275	.085
19	−.224	.073
20	−.103	.076
Literacy	−.076	.036
Date	−.0093	.0024
Antigua[c]	−.260	.235
Barbados	−.005	.268
Jamaica	−.084	.209
Other West Indies	−.363	.221
Maryland	.194	.063
Other mainland	−.039	.072
February[d]	−.100	.092
March	.171	.113
April	.156	.119
May	−.095	.164
June	−.431	.132
July	−.196	.113
August	−.486	.098
September	−.223	.096
October	−.400	.101
November	−.225	.100
December	.014	.097
Sugar[e]	−.0162	.0074
Constant	4.830	. . .
R^2	.566	. . .
F	52.06	. . .
n	2,044	. . .

SOURCE.—London, Corporation of London Records Office; Kaminkow and Kaminkow (1964); Galenson (1977*a*).

NOTES.—See table 1 for variables not defined here.

[a] Female age interactions: indicated variable = 1 for women of given age.

[b] Trade age interactions: indicated variable = 1 for men of given age who recorded a trade (as defined in table 1).

[c] For destinations, zero class = Virginia.

[d] For months, zero class = January. Separate interaction terms between West Indian destination and month of registration were included in the equation, but their coefficients were generally small in value and are not reported.

[e] Sugar = average annual price of muscovado sugar in London, in shillings per hundredweight, lagged 1 year, for servants bound for West Indies (Sheridan [1974, pp. 496–97], with linear interpolation for 1717–20 and 1727).

difference in the length of term by sex. In a suggestive parallel result, Robert Fogel and Stanley Engerman found that, excluding the value of childbearing, the net earnings of female slaves were greater than those of men prior to the age of 18, apparently due to the more rapid physical maturation of women.[20] The sex differentials in the terms of young indentured servants might have resulted from the same source.

The results of table 1 indicate that premia for skills were reflected in the length of servants' indentures. The more detailed specification of the occupations presented in table 1 for the London sample indicates that the marginal premium paid for servants in each of four occupational categories—farmers, services, metal and construction crafts, and clothing and textile trades—was virtually the same.[21] Unskilled laborers received terms longer than servants with skilled occupations but shorter than those with no recorded occupations;[22] the latter result may indicate that some premium was paid for the laborers' work experience.

Economists have devoted considerable attention to the analysis of the relationship between productivity and age and have accumulated much evidence on the association between age and wages in recent periods. Less is known of the nature of this relation in past times. It is, therefore, of some interest to consider in more detail the implications of the estimated relationships between age and length of indenture for the age-earnings profiles of servants.[23] Table 3 presents estimates of the relative annual net earnings of servants by age for unskilled and skilled men. The calculations are based on the assumption that the expected colonial sale price of each individual's contract was equal to the constant marginal cost of delivering servants to the colonies.[24] The relationship between age and net earnings is assumed to have been linear, making average net productivity during the term equal to net productivity at the term's midpoint. The estimates of average net

[20] Fogel and Engerman 1974, p. 77; see also Metzer 1975, pp. 136–37. Interestingly, evidence on the hourly earnings of North Carolina cotton mill employees in 1907 indicates that average female earnings were above those of males through the age of 15, equal at 16, and below male earnings thereafter (Wright 1980, p. 6).

[21] None of the four coefficients is significantly different from any of the other three at the .10 level. On the value of skilled servants in the colonies, see, e.g., Martin (1761, vol. 4, fol. 97, verso); Jeaffreson (1878, 1:186); and Galenson (1979*a*, pp. 314–19).

[22] The coefficient of laborer for the London sample in table 1 is significantly different from that of farmer at the .10 level for a one-tailed *t*-test, from those of metal-construction and clothing-textiles at .05, and from that of services at .025.

[23] It might be noted that the erratic behavior of the coefficients of higher ages in table 1 for the Middlesex sample may have been due to the falsification of the ages of some servants. This may have resulted from the legal requirements under which the registrations were made; for discussion and evidence see Galenson (1979*a*, appendix to chap. 3).

[24] Possible seasonal variation in delivery costs has been controlled for in the equation reported in table 2.

productivity are derived from the following formulation of the mean present value of the contracts of servants of age j at the beginning of the term:

$$PV_j = \sum_{i=1}^{n} \frac{NP_i - w_i}{(1 + r)^i},$$

where NP = expected mean annual net productivity during the term; n = mean length of term for servants in each entering age group; w = mean annual wage payments made to servants during the term; r = discount rate.

To solve for the value of NP for each entering cohort, the mean present value of the contracts was set equal to £10, an estimate of the marginal cost of delivering servants to the colonies. The mean age of each group at the time of binding was taken as the recorded age plus one-half year to allow for the rounding of age upon registration. The mean length of term by age was derived from the coefficients of table 2, while the mean annual wage payments were taken directly from the indenture contracts.[25] The estimates were made with a discount rate of 10 percent.

The estimates of table 3, which indicate that the net-earnings profile of skilled servants was steeper than that of the unskilled, are consistent with the normal positive relationship between the steepness of age-earnings profiles and training. The ages at which the servants considered here were bound, between 15 and 20, were prime ones for training in a wide variety of skilled crafts in preindustrial England, through either apprenticeship or less formal arrangements. It is, therefore, not surprising that the net productivity of those in skilled trades rose rapidly during this period of the life cycle.

Skilled servants received a considerable premium: Even at age 15, a

[25] No wages were recorded for unskilled servants. The average wage payments received by the skilled servants in the London sample by age were as in the table below. The present value of freedom dues at the time of binding varied across colonies and over time, according to differences and changes in legislation and with changes in the value of colonial currencies and commodities. No explicit allowance has been made for the dues in the calculation because of the difficulty of estimating their typical value; inclusion of the effect of the lump-sum payment would lower the estimates of the net annual earnings of servants without changing their relative values by age.

Age (Years)	Mean Annual Wage (£)
15	.0
16	.40
17	.79
18	1.45
19	1.52
20	3.02

TABLE 3

ESTIMATED RELATIVE NET ANNUAL EARNINGS OF SERVANTS BY AGE

Unskilled			Skilled		
Age	Mean Net Annual Earnings (£)	Relative Net Earnings (Age 22.9 = 1.00)	Age	Mean Net Annual Earnings (£)	Relative Net Earnings (Age 22.9 = 1.00)
19.1	2.01	.742	18.4	2.39	.413
19.7	2.20	.812	19.2	2.86	.495
20.4	2.36	.871	20.1	3.32	.574
21.2	2.52	.930	21.0	4.08	.706
22.0	2.60	.959	21.9	4.22	.730
22.9	2.71	1.000	22.9	5.78	1.000

NOTE.—Calculated from table 2 and n. 24. See text for procedure. The unskilled profile is calculated from the basic age profile of table 2, that of the skilled from the basic age profile combined with the skilled ("trade") coefficients.

skilled servant received a term 21 percent shorter than his unskilled counterpart. The existence of a sizable premium at such an early age could have been due in part to differences in the relative average work experience of the skilled and unskilled. Thus, possession of a skilled trade at any age implied prior work experience. The age of entry into the labor force for those men registered without occupations cannot be determined, but it is possible that the typical age of entry for the unskilled into employment was that at which English boys normally left home to live in service, roughly 15. If this were the case, work experience and acquired on-the-job training might have accounted for a significant portion of the premium for skilled workers. This would particularly be true for the younger servants, as the relative level of work experience of a skilled to an unskilled worker would be greatest at the lower ages observed here and would decline with age thereafter. That the ratio of skilled to unskilled net earnings increased with age would appear to be strong evidence of the presence of formal training for those in the skilled group.

An interesting feature of the relative age–net-earnings profile of the unskilled shown in table 3 is its close resemblance to those profiles obtained by Fogel and Engerman for unskilled male slaves in the southern United States during 1790–1860 (1972, charts 3 and 4). In view of the considerable differences among these samples with respect to such variables as location and crops cultivated, the similarity of the shapes of the profiles might suggest the importance of physiological factors, particularly the rate of physical maturation, in determining the change, with age, in the productivity of unskilled workers

under conditions of plantation agriculture in the eighteenth and nineteenth centuries.

The relatively small premium paid for the ability to sign in the later sample may have been due both to the abundance of literate servants and to high literacy rates in the colonies. The decline in the size of the reduction of the term due to the ability to sign between the dates of the two samples may have resulted in part from a considerable increase in literacy among the servants, as only 35 percent signed in the earlier sample compared with 67 percent in the later one.[26]

In both samples, with other characteristics constant, servants bound for the West Indies tended to receive shorter terms than those bound for mainland colonies. That servants who immigrated to the West Indies received shorter terms to compensate them for their choice is consistent with the fact that, while both working conditions for servants and economic opportunities for freedmen were known to be poor in the islands after the introduction of large-scale sugar cultivation in the second half of the seventeenth century—with its attendant slave gangs and consolidation of small farms into large estates—the mainland long continued to be considered a land of opportunity for poor immigrants, where freed servants could hope to own land and become prosperous members of society. A persistent theme of West Indian complaints appeared in a 1675 petition sent to the king of England by the Council and Assembly of Barbados: "In former tymes Wee were plentifully furnished with Christian [i.e., white] servants from England . . . but now Wee can gett few English, having noe Lands to give them at the end of their tyme, which formerly was their main allurement." The higher mortality rates of the West Indian colonies decreased servants' expected productivity and made them reluctant to go to the region. However, in conjunction with the high productivity of labor in sugar production, those rates acted to raise the demand for new flows of replacement immigrant labor. That servants bound for the West Indies received terms shorter than those bound for the mainland, in spite of the higher mortality rates in the islands, implies that the marginal productivity of labor was higher in the West Indies than in the mainland colonies.[27]

Both the lower estimated intercept for the later sample and the estimated negative time trend of the later sample indicate a secular decline in the length of indenture. The direction of change is consis-

[26] On the relationship between ability to sign and other aspects of literacy in this period, see Schofield (1968, pp. 311–25). The difference between the coefficients of literacy in the two samples is significant at .10 for a two-tailed t-test. On the servants' ability to sign, see Galenson (1979*b*).

[27] London, Public Record Office 1675. On relative mortality rates in the West Indian and mainland colonies, see Wells (1975, pp. 280–82).

tent with a number of long-term trends, including rising reservation wages of servants due to rising real wages in England between the mid-1680s and the middle of the eighteenth century, falling real shipping costs, and declining colonial mortality rates which could have produced a secular increase in the colonial demand for labor.[28] The negative estimated trend further suggests the presence of a secular increase in real wages in the colonies during the period spanned by these two samples.

A consistent seasonal pattern in the length of indenture appears in the results of table 2, as servants bound for mainland destinations whose indentures were signed between June and November received sizable reductions in their terms relative to servants bound in winter and spring. Most of these servants were bound for Maryland or Virginia, and the observed pattern could have been due to the effect of the seasonality of tobacco production on the costs involved in supplying servant labor to the Chesapeake. The shipping patterns which resulted from the timing of the harvests dictated that the amount of backhaul space for servants was greatest in summer and early fall. Since servants were provided with food and lodging from the time they signed their contracts, the cost of delivering a servant to the colonies may have declined in peak shipping seasons because the more frequent departures of ships reduced the average waiting time in port between binding and sailing. The lower costs of the peak seasons could, therefore, have resulted in shorter indentures for servants bound in peak seasons than for those bound in slack shipping months.[29]

Another potential source of variation in the length of indenture was annual changes in the colonial demand for labor. The results presented in table 2 indicate that the lagged annual average price of muscovado sugar in London had a significant and negative effect on the length of indentures of servants bound for the West Indies during 1718–59; the estimated effect of a change in the price of sugar from its minimum to its maximum in the period, with other things equal,

[28] On English wages, see Gilboy (1934, pp. 219–25) and Phelps Brown and Hopkins (1956, pp. 302–13). On changes in colonial mortality rates, see, e.g., Menard (1977*a*, pp. 99–100). The organization and quality of information in the market for contracts may have improved during the period spanned by the two samples. It is suggestive that the coefficient of variation of the term of indenture among men of a given age was considerably lower in the later than the earlier sample for six of the eight age groups of minors above the age of 12. On the relation of wage dispersion to information, see Stigler (1962).

[29] On the seasonality of tobacco production and shipping, see Bullock (1649, p. 46); Alsop (1666, p. 51); and Bruce (1907, vol. 1, pp. 622–24). On the costs of maintaining servants between binding and sailing, see Smith (1947, pp. 36–37, 59–65) and Scottish Record Office (GD 23/6/98, nos. 4, 14, 18). For factors relevant to seasonality in the colonial demand for labor, see Mullin (1972, p. 15) and Morgan (1975, p. 158).

was a reduction of about 5 months in the term of indenture. The sign of this effect would be the one predicted if, as appears to have been the case, high sugar prices normally resulted from high levels of demand for sugar rather than reductions in supply, for high sugar prices would then have tended to produce high demand for labor and, ceteris paribus, to shorten terms.[30]

IV. Conclusions

The price paid for human capital in the colonial American market for indentured servants varied systematically with respect to factors which influenced servants' productivity, as economic theory predicts. All servants incurred debts of similar value in immigrating to the colonies and sold claims on their future labor, in the form of indentures, to repay these debts. Characteristics which raised the expected productivity of servants in the American colonies raised the market valuation of their human capital and, therefore, shortened the term for which the servant was bound. Thus, servants with skilled trades and those able to sign served shorter terms than the unskilled and illiterate of similar age and sex. Women were found to have received shorter terms than men until the age of 18, perhaps due to their earlier maturation. The results also indicated that servants bound for the West Indies received shorter terms in compensation for their undesirable destinations. Servants bound during peak shipping seasons were found to have received reduced terms, perhaps due to the shorter average waiting time prior to departure in those months when backhaul cargo space was most abundant. Finally, the length of indenture of servants bound for the West Indies was found to have varied inversely with the price of sugar, suggesting that increases in the colonial demand for labor shortened the term of servitude.

Among the issues which need further investigation are the precise reasons for the observed shapes of the age-earnings profiles of servants. Additional research, including the collection of evidence on wage rates in the colonies by age and skill, may serve to distinguish

[30] Similar analysis of the contracts of servants bound for the Chesapeake during 1718–40 indicates that the lagged annual farm price of Maryland tobacco had no significant effect on the length of indenture; in a regression equation which included the variables of table 2, the estimated coefficient of an interaction term between the price of tobacco, lagged 1 year, and Chesapeake destination was insignificant. Due to a greater continuing reliance on production of a single staple in the West Indies in the eighteenth century, the price of sugar may serve as a better index of the West Indian demand for labor than does the price of tobacco for the Chesapeake. On diversification of agricultural production in the Chesapeake, see Clemens (1974, pp. 100–148) and Stiverson (1977, pp. 65–103).

and isolate the effects of such contributory factors as physical maturation and investment in human capital in producing the age-earnings relationships which lay behind the market valuation of indentured human capital. What the present research has indicated is that the application of economic analysis to quantitative evidence generated by the system of indentured servitude can provide information on the way in which the market once explicitly evaluated stocks of human capital and, in so doing, can yield new insights into the operation of labor markets in early America.

Appendix

For the Middlesex sample of 1683–84, a regression equation was estimated with the amount of the cash payment (in shillings, sterling) made to a servant as the dependent variable, with the same independent variables used in the analysis of the length of indenture. The sample used was that of all servants, minors and adults, whose contracts contained all the necessary information. The results are as shown in table A1.

The hypothesis that all the coefficients are simultaneously equal to zero can be rejected at the .01 level, but the proportion of the variance explained is small, as might be expected in view of the rarity of the payments. Only one coefficient is significant at the .01 level: Possession of a skilled trade, which table 1 shows to have reduced the term of an indenture, significantly raised the expected cash payment made to a servant.

As noted in the text, cash payments appear to have been recorded only for minors in the London sample of 1718–59. The form of the payments recorded differs from that of the earlier sample: Whereas the cash payments to servants in the Middlesex contracts of 1683–84 appear to have been simple lump-sum payments made at the time of binding, those contracted for in London in the eighteenth-century sample were generally salaries to be paid

TABLE A1

Estimated Regression Coefficients, Middlesex Sample, 1683–84

Independent Variable	Estimated Coefficient	Standard Error
Age (years)	.011	.038
Sex[a]	−.485	.456
Literacy[b]	−.443	.363
Trade[c]	1.466	.399
West Indies[d]	.088	.354
Constant	.323	...
R^2	.026	...
F	4.03	...
n	760	...

Source.—London. Greater London Record Office (Middlesex Section).
[a]Male = 0, female = 1.
[b]Marked = 0, signed = 1.
[c]Laborer or no occupation = 0; all other occupations = 1.
[d]West Indian destinations = 1; mainland = 0.

TABLE A2

Estimated Regression Coefficients, London Sample, 1718–59

Independent Variable	Estimated Coefficient	Standard Error
Age (years)	.144	.054
Sex	.306	.329
Literacy	.428	.167
Trade	1.397	.190
West Indies	1.291	.165
Constant	−3.141	...
R^2	.094	...
F	42.20	...
n	2,046	...

Source.—London, Corporation of London Records Office.
Note.—All variables defined as in table A1.

annually in local currency. A regression was estimated with the amount of the annual salary (in pounds local currency) as the dependent variable, with the same independent variables used above. (The dependent variable is the unadjusted value of the colonial currency. Although most of the colonies' currencies were devalued relative to sterling, the differences in currency values across the principal American colonies in most of this period were small. See McCusker 1978.) The sample used was that of all minors (age less than 21) whose contracts contained all the necessary information. The results were as shown in table A2.

The relationship is again statistically significant, and the proportion of the variance explained is again low. Age, literacy, possession of a skilled trade, and West Indian destinations are all significant at .01 and positively related to the servant's salary; in table 2 it is shown that all are significantly and negatively associated with the length of indenture.

The results of the analysis of servants' cash payments and salaries in both samples are consistent with the hypothesis that these were positively related to the servants' expected earnings in the colonies. These results therefore reinforce the analysis presented of the length of indenture.

References

Alexander, Edward P., ed. *The Journal of John Fontaine.* Charlottesville: Univ. Press Virginia, 1972.

Alsop, George. *A Character of the Province of Mary-Land.* London: Peter Dring, 1666.

Annapolis. Maryland Hall of Records. Maryland Probate Records.

Becker, Gary S. *Human Capital.* 2d ed. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1975.

Bruce, Phillip Alexander. *Economic History of Virginia in the Seventeenth Century.* New York: Macmillan, 1907.

Bullock, William. *Virginia Impartially Examined, and Left to Publick View, to Be Considered by All Iudicious and Honest Men.* London: John Hammond, 1649.

Clemens, Paul G. E. "From Tobacco to Grain: Economic Development on

Maryland's Eastern Shore, 1660–1750." Ph.D. dissertation, Univ. Wisconsin, 1974.

Dunn, Richard S. *Sugar and Slaves: The Rise of the Planter Class in the English West Indies, 1624–1713.* New York: Norton, 1973.

Edinburgh. Scottish Record Office. Bught Papers. GD 23/6/98.

Erickson, Charlotte. *American Industry and the European Immigrant, 1860–1885.* Cambridge, Mass.: Harvard Univ. Press, 1957.

Fogel, Robert William, and Engerman, Stanley L. "The Market Evaluation of Human Capital: The Case of Slavery." Paper presented at the Cliometrics Conference, Madison, Wis., 1972.

———. *Time on the Cross: The Economics of American Negro Slavery.* Boston: Little, Brown, 1974.

Galenson, David W. "Agreements to Serve in America and the West Indies, 1727–31." *Genealogists' Magazine* 19 (June 1977): 40–44. (*a*).

———. "Immigration and the Colonial Labor System: An Analysis of the Length of Indenture." *Explorations Econ. Hist.* 14 (October 1977): 360–77. (*b*).

———. "The Indenture System and the Colonial Labor Market: An Economic History of White Servitude in British America." Ph.D. dissertation, Harvard Univ., 1979. (*a*).

———. "Literacy and the Social Origins of Some Early Americans." *Historical J.* 22 (March 1979): 75–91. (*b*).

Geiser, Karl Frederick. *Redemptioners and Indentured Servants in the Colony and Commonwealth of Pennsylvania.* New Haven, Conn.: Tuttle, Morehouse & Taylor, 1901.

Gilboy, Elizabeth W. *Wages in Eighteenth Century England.* Cambridge, Mass.: Harvard Univ. Press, 1934.

Gray, Lewis Cecil. *History of Agriculture in the Southern United States to 1860.* Gloucester, Mass.: Peter Smith, 1958.

Greene, Evarts B., and Harrington, Virginia D. *American Population before the Federal Census of 1790.* New York: Columbia Univ. Press, 1932.

Greene, Lorenzo Johnston. *The Negro in Colonial New England, 1620–1776.* New York: Columbia Univ. Press, 1942.

Heavner, Robert O. "Economic Aspects of Indentured Servitude in Colonial Pennsylvania." Ph.D. dissertation, Stanford Univ., 1976.

Jeaffreson, John Cordy, ed. *A Young Squire of the Seventeenth Century. From the Papers (A.D. 1676–1686) of Christopher Jeaffreson.* London: Hurst & Blackett, 1878.

Kaminkow, Jack, and Kaminkow, Marion. *A List of Emigrants from England to America, 1718–1759.* Baltimore: Magna Charta, 1964.

Kingsbury, Susan M., ed. *The Records of the Virginia Company of London.* Washington: Government Printing Office, 1906.

Littleton, Edward. *The Groans of the Plantations.* London: M. Clark, 1689.

London. Corporation of London Records Office. "Memoranda of Agreements to Serve in America and the West Indies."

London. Greater London Record Office (Middlesex Section). "Plantation Indentures." MR/E.

London. Public Record Office. Colonial Office, class 1, piece 35, f. 237v. "Petition of the Council and Assembly of Barbados to the King." 1675.

———. High Court of the Admiralty, class 20, piece 636. "Ledger for Goodes Sould in ye Barbadoes . . . Sent by ye Shipp Abraham 1636. . . ." 1636.

———. High Court of the Admiralty, class 20, piece 635. "Booke of Accompte

for the Shippe Called ye Tristam and Jeane of London wch Came from Virginia Anno Dm 1637." 1637.

McCusker, John J. *Money and Exchange in Europe and America, 1600–1775: A Handbook.* Chapel Hill: Univ. North Carolina Press, 1978.

Main, Gloria Lund. "Personal Wealth in Colonial America: Explorations in the Use of Probate Records from Maryland and Massachusetts, 1650–1720." Ph.D. dissertation, Columbia Univ., 1972.

Martin, Samuel. "Letter Book of Samuel Martin Senior." British Museum, Martin Papers, Add. MSS 41,349, 1761.

Maryland Historical Society. *The Calvert Papers.* Baltimore: Maryland Hist. Soc., 1889.

Menard, Russell R. "Economy and Society in Early Colonial Maryland." Ph.D. dissertation, Univ. Iowa, 1975.

———. "Immigrants and Their Increase: The Process of Population Growth in Early Colonial Maryland." In *Law, Society, and Politics in Early Maryland,* edited by Aubrey C. Land, Lois G. Carr, and Edward C. Papenfuse. Baltimore: Johns Hopkins Univ. Press, 1977. (*a*)

———. "From Servants to Slaves: The Transformation of the Chesapeake Labor System." *Southern Studies* 16 (Winter 1977): 355–90. (*b*)

Metzer, Jacob. "Rational Management, Modern Business Practices, and Economies of Scale in the Ante-bellum Southern Plantations." *Explorations Econ. Hist.* 12 (April 1975): 123–50.

Middleton, Arthur Pierce. *Tobacco Coast: A Maritime History of Chesapeake Bay in the Colonial Era.* Newport News, Va.: Mariners' Museum, 1953.

Morgan, Edmund S. *American Slavery, American Freedom: The Ordeal of Colonial Virginia.* New York: Norton, 1975.

Mullin, Gerald W. *Flight and Rebellion: Slave Resistance in Eighteenth-Century Virginia.* London: Oxford Univ. Press, 1972.

Nicholson, Cregoe D. P. *Some Early Emigrants to America.* Baltimore: Genealogical Pub. Co., 1965.

Pares, Richard. *Merchants and Planters. Econ. Hist. Rev.,* suppl. no. 4. Cambridge: Cambridge Univ. Press, 1960.

Phelps Brown, E. H., and Hopkins, Sheila V. "Seven Centuries of the Price of Consumables, Compared with Builders' Wage-Rates." *Economica* 23 (November 1956): 296–314.

Purchas, Samuel. *Purchas His Pilgrimes.* London: Henrie Featherstone, 1625.

Rosen, Sherwin. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *J.P.E.* 82, no. 1 (January/February 1974): 34–55.

Schofield, R. S. "The Measurement of Literacy in Pre-industrial England." In *Literacy in Traditional Societies,* edited by Jack Goody. Cambridge: Cambridge Univ. Press, 1968.

Semmes, Raphael. *Crime and Punishment in Early Maryland.* Baltimore: Johns Hopkins Univ. Press, 1938.

Sheridan, Richard B. *Sugar and Slavery: An Economic History of the British West Indies, 1623–1775.* Barbados: Caribbean Univ. Press, 1974.

Smith, Abbot Emerson. *Colonists in Bondage: White Servitude and Convict Labor in America, 1607–1776.* Chapel Hill: Univ. North Carolina Press, 1947.

Smith, John. *The Generall Historie of Virginia, New-England, and the Summer Isles.* London: Michael Sparkes, 1624.

Stigler, George J. "Information in the Labor Market." *J.P.E.* 70, no. 5, pt. 2 (October 1962): 94–105.

Stiverson, Gregory A. *Poverty in a Land of Plenty: Tenancy in Eighteenth-Century Maryland.* Baltimore: Johns Hopkins Univ. Press, 1977.

Sutherland, Stella H. *Population Distribution in Colonial America.* New York: Columbia Univ. Press, 1936.

Taunton. Somerset Record Office. Helyar Manuscripts.

Tullideph, Walter. "Letter Books of Dr. Walter Tullideph." Unpublished. Dundee, Scotland, 1759.

Wareing, John. "Some Early Emigrants to America, 1683–84: A Supplementary List." *Genealogists' Magazine* 18 (March 1976): 239–46.

Wells, Robert V. *The Population of the British Colonies in America before 1776.* Princeton, N.J.: Princeton Univ. Press, 1975.

Wilson, Samuel. *An Account of the Province of Carolina in America.* London: Francis Smith, 1682.

Wood, Peter H. *Black Majority: Negroes in Colonial South Carolina from 1670 through the Stono Rebellion.* New York: Norton, 1975.

Wright, Gavin. "Cheap Labor and Southern Textiles, 1880–1930." Paper presented at the Workshop in Economic History. Univ. Chicago, January 25, 1980.

The Welfare Cost of Capital Income Taxation in a Growing Economy

Christophe Chamley

Yale University

The welfare cost of capital income taxation is analyzed in a general equilibrium framework, where the private sector is represented by a competitive household endowed with perfect foresight and an infinite life. The value of the welfare cost depends essentially on the elasticity of substitution between capital and labor in the production function. Numerical estimates are presented for different values of the parameters of the model. The welfare gain obtained by the abolition of the capital income tax is smaller when the private sector is not endowed with perfect foresight (it is reduced by about 40 percent when expectations are myopic). The allocation efficiency cost of the corporate tax dwarfs the intertemporal welfare cost.

A central issue in the current debate on tax reform is the efficiency cost of a tax on capital income. Such a tax introduces a wedge between the prices of consumption at different dates and distorts the intertemporal allocation of resources.

Previous studies on the efficiency cost of the capital income tax have relied on simplifying assumptions which rule out important consequences of capital taxation. In general, these studies are of two types. The first (Krzyzaniak 1967; Sato 1967; Feldstein 1974*a*, 1974*b*; Friedländer and Vandendorpe 1978) analyzes the capital income tax in the context of a neoclassical growth model with endogenous capital stock and factor prices but does not allow an optimal response of

I am grateful to the Cowles Foundation for financial support. Comments by Laurence Kotlikoff, Dale Jorgenson, Laurence Weiss, John Shoven, and anonymous referees were very helpful. This paper would not have been completed without the collaboration of William Brainard.

[*Journal of Political Economy*, 1981, vol. 89, no. 3]

household to changes of taxes. Typically the levels of consumption and labor supply do not depend on future prices and are not consistent with utility maximization. This method leads to results about the long-run incidence of taxation on aggregate variables; the transition to the steady state can be examined by numerical simulations. The second type of study (Levhari and Sheshinski 1972; Feldstein 1978) allows for maximizing behavior but assumes that the factor prices are exogenous and independent of the capital accumulation. The welfare cost of the capital income tax is measured by an application of the Harberger-Hicks-Hotelling formula (Green and Sheshinski 1979) to the intertemporal framework.

The welfare cost of the capital income tax is analyzed here in a stylized intertemporal general equilibrium model (described in the next section). Issues about intra- or intergenerational equity are ignored. Therefore, it is assumed that the private sector can be represented by a household with an infinite life.[1]

This household determines the level of consumption by the maximization of its intertemporal utility function. Future factor prices (wage and interest rates) depend on the accumulation of capital through a neoclassical technology and are known with perfect foresight. The household behaves competitively: Future (endogenous) prices are taken as given.

The welfare cost of the capital income tax is analyzed in the second section. Following common practice, we consider the welfare cost induced by a tax with lump-sum redistribution. Initially, the economy is assumed to be on the balanced growth path where capital income is taxed at a fixed rate (and tax revenues are refunded). At time zero, the tax is abolished (together with the refunds). Thereafter, the economy moves on a dynamic path toward a new steady state. The welfare cost of the capital income tax is equal to the welfare gain obtained by the abolition of the tax, namely, by the difference between the level of utility on the new dynamic path (after the tax reform) and the level of utility on the initial balanced growth path (with the tax in effect).[2] As usual, the welfare cost of the tax is

[1] It would be sufficient to assume that individuals' utilities depend on their consumption in their own finite lifetime and on the welfare of their immediate descendants (for a development of this argument, see Barro [1974]). If individuals do not leave a bequest, the problem of intergenerational equity arises, and the concept of excess burden used here does not apply: Even in the case where no revenue has to be raised, the no-tax solution is not the first best; the level of capital has no optimal property (Diamond 1965). In order to redistribute income among generations, a benevolent social planner would impose a tax or a subsidy on capital income (and possibly on other goods). For an analysis of taxation in a general equilibrium model where individuals maximize a lifetime utility function with no bequest, see Hall (1969), Diamond (1970), Summers (1979), and Chamley (1980*b*).

[2] We could also consider the imposition of a tax on the initial balanced growth path

measured by a wealth equivalent and is of a second order with respect to the tax rate. A second approximation of this excess burden is given, which depends on the parameters of the utility and production functions and on the growth rate. An extension of the Levhari and Sheshinski (1972) result is obtained as a special case. Since the general excess-burden formula is exact only for infinitesimal values of the tax rate, the error of this second-order approximation is analyzed in a numerical example.

The assumption of perfect foresight is relaxed in Section III. Because the dynamic path after the abolition of the tax is no longer optimal, the welfare gain induced by the tax reform is smaller in this case. The case of myopic expectations is an important example of the more general class of expectations which are considered. In the following sections we revert to the assumption of perfect foresight.

The case where the tax rate on capital income is not identical for all sectors of production (an example is found in the corporation tax) is considered in Section IV. The intertemporal welfare cost of the tax is compared with the inefficiency cost due to the misallocation of capital between the different sectors of production.

In Section V, the assumption of a fixed labor supply is relaxed. Since the capital income tax lowers the long-run wage rate, its excess burden depends on the (compensated) elasticity of the labor supply.

This analysis of the capital income tax relies on a stylized model.[3] However, numerous numerical examples show that some of the results obtained are fairly robust. In the conclusion, these results are summarized and related to other studies using more disaggregated models.

I. The Model

There is one good in the economy. This good can be consumed or used as capital in the production process. Total output per efficiency unit of labor, net of capital depreciation, is given by the neoclassical production function, $y = f(k)$, where k is the level of the capital stock per efficiency unit of labor.

The private sector is represented by a household, growing at the rate n, which takes prices as given and maximizes under its budget constraint the utility function,

$$U = \int_0^\infty e^{-\rho t} e^{nt} u(c_t e^{\mu t}) dt, \tag{1}$$

with no taxation. However, the analysis of the abolition (instead of the imposition) of the tax is technically more simple and suits better the problems of tax reform. It has been verified that for small tax rates, the two methods give the same results.

[3] The same method could have been used to analyze the excess burden of the labor income tax or the income tax (Chamley 1980a).

where the following notation is used: ρ = pure rate of time preference, μ = rate of growth of labor augmenting technological change, and c_t = consumption per unit of efficient labor.

The function u will be assumed to be of the form $u(c) = c^{1-\sigma}$.[4] The labor supply per capita is fixed and normalized to one at time zero. The representative household is endowed with perfect foresight and behaves competitively, taking the endogenous future prices (wage and interest rates) as given.

Because of the first-order condition in the maximization of the utility function, the dynamic path of the economy satisfies the equation[5]

$$\dot{c}_t = \frac{c_t}{\sigma}(r_t - \rho^*), \tag{2}$$

where r_t is the net rate of return available to the household, and $\rho^* = \rho + \sigma \cdot \mu$.

The capital accumulation is defined by

$$\dot{k}_t = f(k_t) - (n + \mu)k_t - c_t, \tag{3}$$

where k_0 represents the initial capital stock.

The dynamic behavior of the economy is defined by equations (2) and (3) and by the initial values k_0 and c_0 at time zero. The initial value of the capital stock k_0 is given. There is a unique value of c_0, such that its associated dynamic path satisfies the budget constraint of the household (which is equivalent here to the transversality conditions). For this value of c_0, the dynamic path converges to the steady state defined by[6]

$$\rho^* = r^* = f'(k^*), \tag{4}$$

$$c^* = f(k^*) - (n + \mu)k^* \tag{5}$$

(an asterisk will denote a steady-state value).

[4] This assumption is necessary for the existence of a competitive balanced growth path when the intertemporal welfare function is additive and the rate of labor augmenting technological change μ is different from zero. (For an introduction to the literature on optimal growth, see Koopmans [1967].) In a discrete-time formulation one could also use a stationary utility function. The existence and stability of optimal balanced growth paths in this context is studied by Iwai (1972).

[5] The consumption levels at time t and $t + \Delta t$ satisfy the relation,

$$\frac{u'[e^{\mu(t+\Delta t)}c_{t+\Delta t}]}{u'(e^{\mu t}c_t)} = \frac{1 + \rho\Delta t}{1 + r\Delta t}.$$

Up to the first order, this expression is equivalent to

$$\frac{u'(e^{\mu t}c_t) + u''(e^{\mu t}c_t)(c_t\mu\Delta t + \Delta c_t)e^{\mu t}}{u'(e^{\mu t}c_t)} = 1 + (\rho - r)\Delta t.$$

A straightforward manipulation gives $(1/c_t)(\Delta c_t/\Delta t) = (1/\sigma)(r - \rho - \sigma\mu)$. When Δt tends to 0, we obtain eq. (2).

[6] The second-order conditions are derived from the concavity of the utility and the production functions. For an exhaustive treatment, see Arrow and Kurz (1970).

The optimal path also defines a consumption function, giving the level of consumption per unit of labor as a function of the capital-labor ratio at each instant (see fig. 1),

$$c = c(k). \tag{6}$$

In the same way, at a given instant, the level of utility is determined by the integral (1) on the optimal path Γ for an initial value of the capital stock which is equal to k: $U = J(k)$.

We now review a few properties of the optimal path which will be useful in the subsequent sections.

The slope of the consumption function $c'(k^*)$ at the stationary point k^* is obtained by taking the limit of the ratio between the relations (2) and (3) when k tends to k^*; $c'(k^*)$ is equal to the positive root of the equation,

$$x^2 - \lambda x - \gamma = 0, \tag{7}$$

where $\lambda = \rho^* - n - \mu$ and $\gamma = -[c(k^*)f''(k^*)]/\sigma = (1/\sigma\epsilon)\big(c(k^*)(r^* + \delta)w^*/\{k^*[f(k^*) + \delta k^*]\}\big)$; ϵ is the elasticity of substitution between capital and labor in the *gross* production function when the capital labor ratio is equal to k^*.[7] Using a first-order approximation of the capital accumulation (3) around k^*, the difference between k and its steady-state value k^* decreases asymptotically at a constant rate a (to a first-order approximation):[8]

$$(k_t \dot{-} k^*) = -a(k_t - k^*), \tag{8}$$

where a is the coefficient of adjustment of the economy toward the steady state and is equal to the difference $c'(k^*) - \lambda$; $-a$ is also equal to the negative root of equation (7).

The same asymptotic rule applies to every endogenous variable z_t in the economy which depends only on the capital-labor ratio (as, e.g., the gross factor prices):[9]

$$(z_t \dot{-} z^*) = -a(z_t - z^*), \tag{8a}$$

where the coefficient a is the same as in equation (8).

[7] The parameter γ can also be expressed as a function of quantities which are easily measurable:

$$\gamma = \frac{(1-\alpha)(r^* + \delta)}{\sigma\epsilon}\left[\frac{r^* + \delta}{\alpha} - (\delta + n + \mu)\right],$$

where α is the share of gross capital income in the production function and δ is the depreciation rate of the capital stock.

[8] This regressive rule may be a very good approximation even if the difference $k - k^*$ is large (see Chamley 1979).

[9] If $z_t = g(k_t)$, to the first order: $(z_t \dot{-} z^*) = g'(k^*)\dot{k}_t = -g'(k^*)a(k_t - k^*) = -a(z_t - z^*)$.

II. The Excess Burden of the Capital Income Tax

Consider now a tax on capital income at the constant rate θ, with lump-sum redistribution of its revenues. The analysis of the model described in the previous section applies with a net interest rate r now given by $r = (1 - \theta)f'(k)$. In particular, in the long run, the net rate of return is constant and still equal to ρ^*. In the long run, the capital income tax increases the gross rate of return and lowers the capital stock and aggregate consumption. The levels of consumption and capital per unit of labor in the steady state with taxation, $\bar{c}$ and $\bar{k}$, respectively, are given by

$$\rho^* = (1 - \theta)f'(\bar{k}) \tag{4a}$$

and

$$\bar{c} = f(\bar{k}) - (n + \mu)\bar{k}. \tag{5a}$$

This steady state is represented by point A in figure 1.

We assume that initially the economy is in the long-run equilibrium with taxation (fig. 1, point A). The deadweight loss of the tax is measured by the welfare gain induced by the suppression of the tax.[10]

At time zero, the capital income tax is abolished[11] (together with the lump-sum refund). The net rate of return suddenly increases, and consumption decreases immediately from $\bar{c}$ to c_0 (the optimal value which depends only on the initial capital stock $k_0 = \bar{k}$, as described in Sec. I). After time zero, because of the increased savings,[12] the capital stock increases. The economy moves toward the new steady state E on the segment BE of the path Γ defined by the dynamic equations (2) and (3) (see fig. 1).

The welfare gain of the tax reform ΔU is given by the difference between the utility on the path BE and the utility in the steady state A:[13]

$$\Delta U = J(\bar{k}) - \frac{u(\bar{c})}{\rho^* - n - \mu}. \tag{9}$$

[10] When the tax revenues are not equal to zero, this is equivalent to the gain obtained by a shift from an interest tax to lump-sum taxation.

[11] It is essential throughout this study that the tax changes are unanticipated; for an analysis of the effects of anticipated tax changes, see Hall (1971).

[12] The observed elasticity of gross savings η, with respect to the net rate of return at time zero (under the assumption of perfect foresight about future interest rates), can be determined from the parameters of the model: $\eta = a\epsilon\{r^*/[(r^* + \delta)(n + \mu + \delta)(1 - \alpha)]\}$, where α and δ are defined in n. 7 and a is given in table 2. For the values used below, the Cobb-Douglas case ($\sigma = \epsilon = 1$), $\eta = .95$.

[13] When the tax rate θ is small, the following experiment is symmetrical and gives the same result: Assume that the economy is in the steady state with no tax (point E). At time zero, the tax is instituted (with lump-sum refunds). The economy moves then on a path toward the point A. Call J_1 the utility level on this path and J_0 the utility level at the steady state E. The excess burden of the interest tax is equal to $J_0 - J_1$.

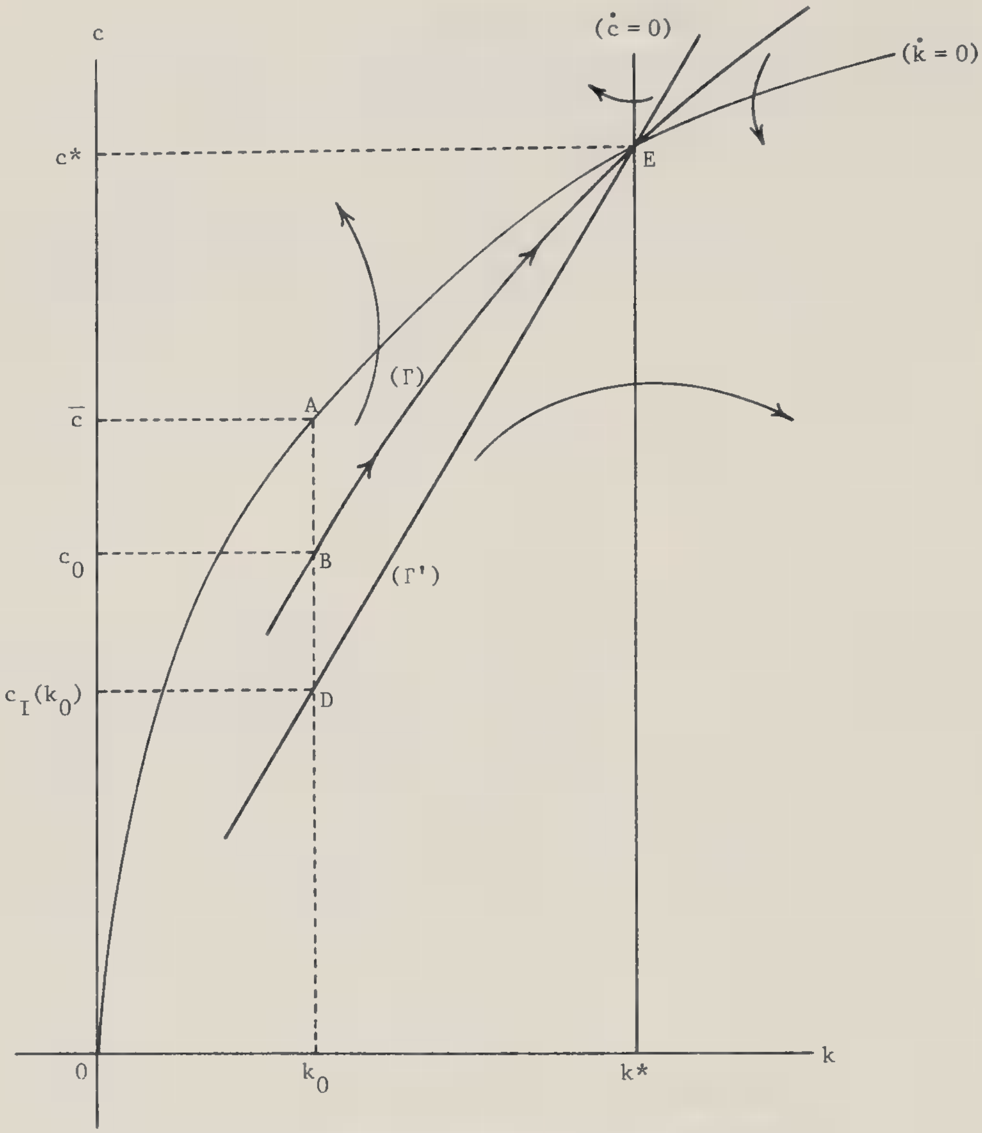


FIG. 1.—The consumption function and the dynamic path

A Taylor expansion of (9) around k^* gives

$$\Delta U \sim J(k^*) + J'(k^*)(\bar{k} - k^*) + \frac{J''(k^*)}{2}(\bar{k} - k^*)^2$$

$$- \frac{u(c^*)}{\rho^* - n - \mu} - \frac{u'(c^*)}{\rho^* - n - \mu}\frac{d\bar{c}}{d\bar{k}}(\bar{k} - k^*)$$

$$- \frac{u''(c^*)}{2(\rho^* - n - \mu)}\frac{d^2\bar{c}}{d\bar{k}^2}(\bar{k} - k^*)^2.$$

From (5a) $d\bar{c}/d\bar{k} = f'(k^*) - n - \mu = \rho^* - n - \mu$. Furthermore, the marginal value of capital $J'(k^*)$ is equal to the marginal utility of

consumption $u'(c^*)$. Therefore,

$$\Delta U \sim \frac{1}{2}\left[J''(k^*) - \frac{u''(c^*)}{\rho^* - n - \mu}\,\frac{d^2\bar{c}}{d\bar{k}^2}\right]\left(\frac{d\bar{k}}{d\theta}\right)^2 \theta^2. \qquad (10)$$

This welfare cost is of a second order with respect to the tax rate. It is convenient to divide it by the marginal utility of consumption at the stationary point E, $u'(c^*)$, in order to obtain its wealth equivalent ΔM:[14]

$$\Delta M = L\theta^2 \frac{c^*}{r^* - n - \mu}, \qquad (11)$$

with

$$L = \frac{1}{2\sigma}\left(\frac{r^* a}{\gamma}\right)^2; \qquad (12)$$

the parameters a and γ have been defined in the previous section.

The relation (11) is to be interpreted as follows: The welfare cost of a capital income tax at the (small) rate θ is equivalent to a permanent reduction of consumption on the balanced growth path by a fraction $L\theta^2$. The variable L depends on the technology, the utility function, and the growth rate. We now consider some of its properties.

The relation (12) can be rewritten,

$$L = L_P\left[\frac{(r^* - n - \mu)a}{\gamma}\right]^2, \qquad (13)$$

where

$$L_P = \frac{1}{2\sigma}\left(\frac{r^*}{r^* - n - \mu}\right)^2. \qquad (14)$$

From (14), and the definition of a and γ (eq. [7]), it is straightforward to derive the following properties:

$$L < L_P;\ \frac{\partial L}{\partial \epsilon} > 0;\ \frac{\partial L}{\partial \sigma} < 0;$$

$$\lim_{\epsilon \to \infty} L = L_P;\ \lim_{\epsilon \to 0} L = 0.$$

The excess burden of the tax is an increasing function of the elasticity of substitution ϵ between capital and labor in the production function. When ϵ is equal to zero, the capital-labor ratio is fixed, and since the labor supply is fixed, there is no distortion. The welfare cost is nil. When ϵ is infinite, the gross factor prices are fixed, $L = L_P$. This

[14] The relations (11) and (12) are derived in the Appendix, Sec. II.

TABLE 1

Normalized Excess Burden of the Capital Income Tax

	ϵ										
σ	.2	.4	.6	.8	1.0	1.2	1.4	1.6	1.8	2.0	∞
.5	1.24	2.41	3.56	4.67	5.76	6.83	7.88	8.91	9.92	10.93	400.0
1.0	1.21	2.33	3.41	4.45	5.46	6.44	7.40	8.33	9.25	10.14	200.0
1.5	1.19	2.28	3.31	4.30	5.25	6.16	7.05	7.92	8.76	9.58	133.33
2.0	1.17	2.23	3.22	4.17	5.07	5.94	6.78	7.59	8.37	9.13	100.0

Note.—Given by eq. (12), in percentage of annual consumption.

TABLE 2

Annual Rate of Convergence toward Balanced Growth Path

	ϵ									
σ	.2	.4	.6	.8	1.0	1.2	1.4	1.6	1.8	2.0
.5	33.98	23.74	19.21	16.51	14.67	13.31	12.25	11.40	10.70	10.10
1.0	23.74	16.51	13.31	11.40	10.10	9.14	8.40	7.80	7.30	6.88
1.5	19.21	13.31	10.70	9.14	8.08	7.30	6.69	6.21	5.80	5.46
2.0	16.51	11.40	9.14	7.80	6.88	6.21	5.68	5.26	4.91	4.62

Note.—a is measured in percentage.

case is analogous to the partial equilibrium situation with exogenous factor prices and provides an upper bound for the general formula.[15]

Since $1/\sigma$ is an index of the intertemporal substitutability of consumption in the utility function, it is not surprising that the excess burden is a decreasing function of σ.

Tables 1 and 2 present estimates of the variable L and of the annual rate of convergence of the economy toward the steady state a for different values of ϵ and σ.[16] The other parameters of the model are chosen to characterize the U.S. economy:[17] $\mu = .02$, $n = .0$, the gross capital income share is equal to .33, the rate of capital depreciation is equal to 5 percent, and ρ^* is equal to the long-run value of the net rate of return and is taken to be equal to 4 percent. There remain only two

[15] This formula has been derived by Levhari and Sheshinski (1972) for a stationary economy. Another proof, using the well-known Harberger-Hicks-Hotelling formula, is given in the Appendix, Sec. I.

[16] These values of the rate of convergence can be compared with those discussed by Sato (1966) and give additional information for a choice of ϵ and σ.

[17] See, e.g., Jorgenson and Christensen 1973.

unknown parameters, ρ and σ. Since there is no general agreement about their values,[18] we present estimates for different values of σ. Once a value is chosen for σ, the pure rate of time preference is implicitly determined by the relation, $\rho = \rho^* - \sigma\mu$.

As can be seen from table 1, for realistic values of ϵ the excess burden is not very sensitive to the elasticity of the marginal utility σ.[19]

The last column in table 1 corresponds to the partial equilibrium case. It is clear that for realistic values of ϵ, the excess burden is much smaller. The variable L is represented as a function of ϵ for a fixed value of σ, equal to one, on figure 2 (curve $\beta = 0$). For the relevant range of ϵ, the excess burden of the interest tax increases almost linearly with ϵ.

The value of the normalized excess burden L depends also on the discount rate r^* and on the growth rate $n + \mu$. No general rule can be derived about the effects of these parameters. When the elasticity ϵ is large, L increases with $n + \mu$ and decreases with r^* (as it is clear from the limit case $\epsilon = \infty$, described by eq. [14]). However, when the elasticity of substitution is smaller than two, and the other parameters of the model have the same values as above, it is found that L is an increasing function of both the discount rate and the growth rate.[20]

The Excess Burden for Large Tax Rates

The measure of the excess burden given by expressions (11) and (12) is exact only for infinitesimal values of the tax rates. In order to estimate the bias involved for large values of θ, we have to use a different method.

At time zero, the economy is on its balanced growth path under taxation (point A on fig. 1), and the initial values of the capital stock and the consumption level are given by (4a) and (5a). The dynamic path after the tax cut, characterized by (2) and (3), is transformed into

[18] Available estimates of σ seem to point to a value higher than one. Weber (1975) reports values between 1.3 and 1.8. Wright's (1969) estimates are somewhat higher—around 4.

[19] In the limit case where the discount rate r^* and the growth rate are equal, the variable L becomes

$$L = -\frac{1}{2}\frac{r^{*2}}{c^*f''(k^*)} = \frac{\alpha}{2(1-\alpha)^2}\cdot\left(\frac{n+\mu}{n+\mu+\delta}\right)^2\cdot\epsilon,$$

which is independent of the utility function and linear in the elasticity ϵ (ceteris paribus).

[20] When $\epsilon = \sigma = 1$, the following values are obtained: Take $n + \mu = 2$ percent; when r^* increases from 2 percent to 6 percent, L increases from 3.00 percent to 6.96 percent. When r^* is equal to 4 percent and $n + \mu$ increases from 2 percent to 4 percent, L increases from 5.46 percent (table 1) to 7.26 percent.

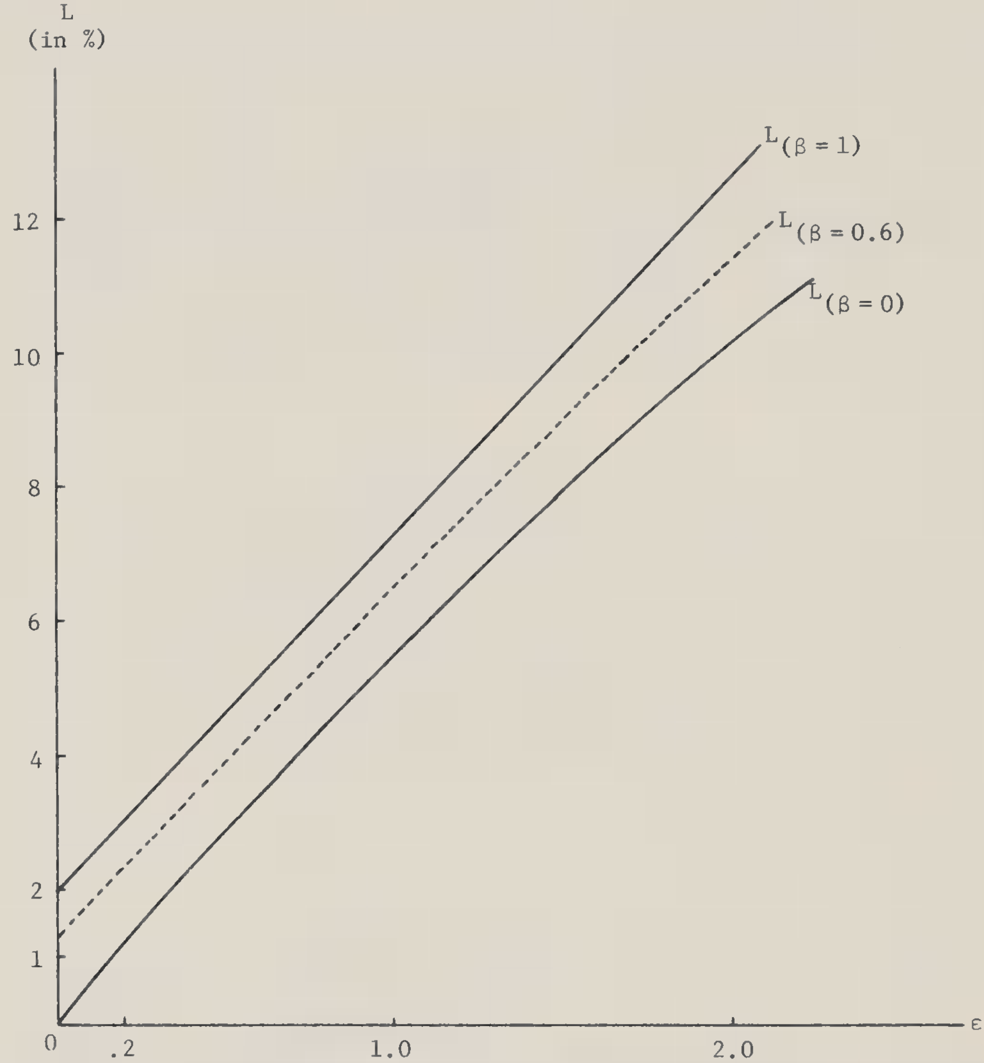


FIG. 2.—The normalized excess burden; ϵ = elasticity of substitution between capital and labor, β = compensated elasticity of substitution between leisure and consumption.

a discrete formulation and simulated numerically using a gradient algorithm of optimal control. The consumption equivalent of the utility after the tax cut c_1 is defined by the stationary level providing the same utility as the dynamic path after the tax cut: $u(c_1)/(\rho^* - n - \mu) = J(\bar{k})$. Table 3 reports the ratio $[(c_1 - \bar{c})/\bar{c}](1/\theta^2)$ as a function of θ for $\sigma = \epsilon = 1$; the other parameters of the model have the same values as in the first two tables.

The first number in the table (for $\theta = 0$) is obtained by formula (12). When θ is small, it gives a good approximation of the normalized excess burden. However, we can see that for large values of the tax rate θ, the approximation formula (12) underestimates the welfare

TABLE 3

NORMALIZED EXCESS BURDEN FOR LARGE TAX RATES (%)

	θ						
	.0	.05	.1	.2	.3	.4	.5
$\frac{c_1 - \bar{c}}{\bar{c}} \cdot \frac{1}{\theta^2}$	5.46	5.82	6.25	7.27	8.58	10.33	12.77

NOTE.—$\sigma = \epsilon = 1$.

gain of the tax cut (for $\theta = 50$ percent, by a factor slightly smaller than $\frac{1}{2}$).[21]

In order to illustrate the previous discussion, let us consider two numerical examples of a shift from capital income taxation to lump-sum taxation.

Assume that the aggregate consumption in the steady state under taxation is normalized to one. When the tax rate θ is equal to 50 percent, total revenues are equal to 12.34 percent. A shift to a lump-sum tax increases the level of welfare by a consumption equivalent of 3.19 percent and the level of consumption in the long run by 8.38 percent. The welfare cost of the capital income tax is equal to 26 percent of the amount of tax revenues. The marginal welfare cost is obtained by multiplying the average cost by a factor of two (this factor should be somewhat greater than two for large rates, since the normalized excess burden increases with the tax rate).

When θ is equal to 30 percent, tax revenues are equal to 6.73 percent. A shift to lump-sum taxation would increase consumption in the long run by 3.3 percent. The welfare cost of the tax is equal to 0.77 percent of the consumption level or 11 percent of the revenues.

The method followed in this section takes into account both the short-run and the long-run incidence of the suppression of the capital income tax. We can observe that the welfare gain of tax reform (measured in consumption equivalent) is much smaller than the percentage increase of consumption in the long run. Also, the excess burden of the capital income tax, although small in terms of aggregate consumption, is not negligible in terms of the revenues generated.

III. The Case of the Nonperfect Foresight

In the previous section, the welfare gain of a reform from capital income taxation to lump-sum taxation was determined under the

[21] On the other side, when θ is large, the model considered here implies a sudden increase of saving which could be too large to be borne by a real economy. Therefore, it is likely that the numbers in table 3 may be overestimated.

assumption that the economy follows an optimal path toward the new steady state. However, when there is no complete set of future markets to convey information on future prices, individuals may not have perfect foresight. In this case, the dynamic path after the tax reform is no longer optimal, and the welfare gain obtained is smaller.[22]

The assumption of perfect foresight is relaxed in this section. For its saving decision, the competitive household relies now on point expectations about the future wage and interest rates. By assumption, these expectations satisfy the following property: At each instant t, anticipated future prices depend only on the value of the capital stock at time t (e.g., through the factor prices at time t) and on some constant parameters (as the long-run values of the factor prices).[23] It follows that the consumption level can be expressed as a function of k, $c_I(k)$.

We also assume that expectations are consistent with the steady state and do not affect its stability. The rate of convergence of the economy a_I depends on the type of expectations considered. For example, in the case of myopic expectations (where the factor prices observed at a given time are expected to be the same in the future), the rate of convergence is greater than in the case of perfect foresight.

As in the previous section the initial situation of the economy is the steady state with taxation. After the suppression of the tax at time zero, the economy converges to the steady state E on the path $\Gamma'(B'E)$, which in general is different from the path Γ. The welfare gain is measured by the difference,

$$\Delta U = \tilde{J}(\bar{k}) - \frac{u(\bar{c})}{r^* - n - \mu}, \tag{15}$$

where $\tilde{J}(\bar{k})$ represents the integral (1) on the path $B'E$.

The second-order approximation of the utility level after the tax cut, $\tilde{J}(\bar{k})$, depends only on the slope of the dynamic path at the steady state,[24] $c_I'(k^*)$, or on the rate of adjustment toward the steady state a_I (since $a_I = c_I'[k^*] - [r^* - n - \mu]$):

$$\tilde{J}(\bar{k}) = F(a_I, \bar{k}); \tag{16}$$

this utility is smaller than the utility on the optimal path,

$$F(a_I, \bar{k}) \leqslant F(a, \bar{k}) = J(\bar{k}). \tag{17}$$

[22] For the class of expectations considered below, the dynamic path still converges to the steady state.

[23] This class includes perfect foresight and myopic, stationary, and regressive expectations. This section uses some intuitive results derived in Chamley (1979).

[24] A proof is given in the Appendix, Sec. III.

The welfare gain induced by the tax cut is equivalent to a permanent increase of the consumption level by a fraction $L_I\theta^2$, where L_I is defined by

$$L_I = [Q(a_I)/Q(a)] \cdot L, \tag{18}$$

with

$$Q(a_I) = \frac{a_I}{\lambda + 2a_I}\left[\frac{a_I}{a \cdot c'(k^*)} - \frac{2}{\lambda}\right]; \tag{19}$$

L is equal to the perfect-foresight value defined by (12) in the previous section.

Using the relations (18) and (19) and the properties of a and $c'(k^*)$ (described in Sec. I), it is straightforward to show that the ratio L_I/L is always smaller than one. When the private sector does not anticipate future prices with perfect foresight, the welfare gain induced by the suppression of the capital income tax is reduced.

The ratio between the values of the excess burden with nonoptimal and optimal adjustments L_I/L is represented in figure 3 as a function of the ratio between the rates of convergence (a_I/a). The different curves correspond to different values of the discount rate r^*. The growth rate $n + \mu$ is equal to 2 percent, and the product $\sigma\epsilon$ is equal to 1.

Also the cases $r^* = 3$ percent and $r^* = 6$ percent provide a very good approximation (less than 1 percent error) of the cases $r^* = 4$ percent, $\sigma\epsilon = \frac{1}{3}$, and $r^* = 4$ percent, $\sigma\epsilon = 2.5$, respectively.

It is interesting to observe that, in general, a shift from capital income taxation to lump-sum taxation always induces a welfare gain (L_I is positive unless a_I has an unrealistically large value, greater than $[12.1] \cdot a$). This gain is close to its maximum for a wide range of values of the ratio (a_I/a). When this ratio is between $\frac{1}{3}$ and 3, L_I is equal to at least 86 percent of L (for $r^* = 4$ percent, $\sigma\epsilon = 1$).

As an example, let us consider the case of myopic expectations. The consumption function takes then the following form:

$$c_S(k) = \left(1 + \frac{1}{\sigma}\frac{\rho^* - r}{r - n - \mu}\right)[f(k) - (n + \mu)k].$$

Using (18) and (19), after elementary manipulations, the ratio between the normalized excess burden under myopic expectations and its value under perfect foresight, L_S/L, is given by

$$L_S/L = \frac{1}{2}\left(1 + \frac{\lambda\sqrt{\lambda^2 + 4\gamma}}{\lambda^2 + 2\gamma}\right), \tag{20}$$

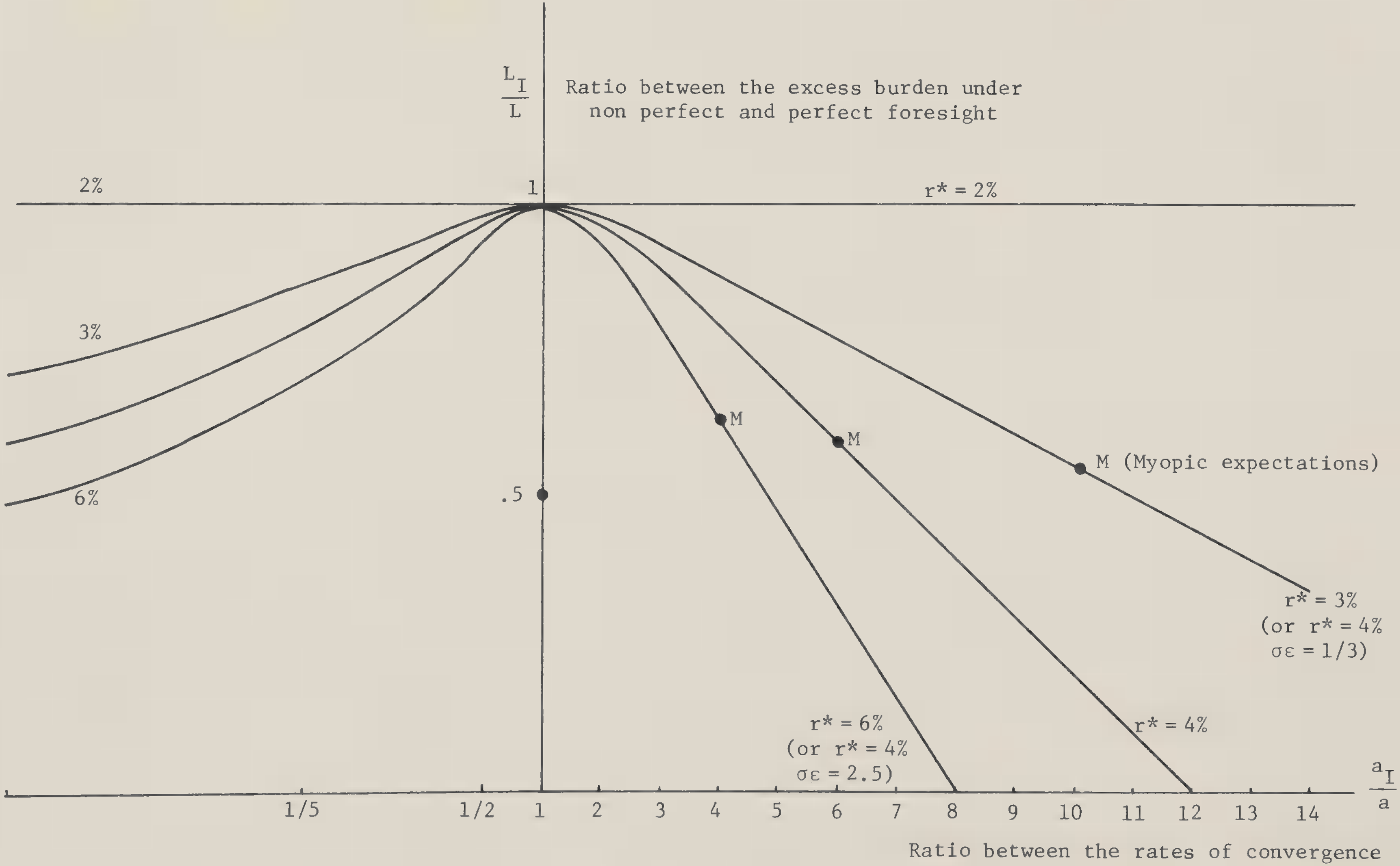


FIG. 3.—The excess burden under nonperfect foresight

with $\lambda = r^* - n - \mu$, and γ is the parameter described in equation (7). A good approximation of (20) is given by

$$L_S/L = \frac{1}{2}\left(1 + \frac{\lambda}{\lambda + a}\right), \tag{21}$$

where a is the optimal rate of convergence reported in table 2.

Under myopic expectations, the welfare gain of the tax reform is always equal to at least 50 percent of the perfect-foresight value. This case is represented by the point M on figure 3 (for $\epsilon = \sigma = 1$, the rate of convergence toward the steady state under myopic expectations a_S is about six times larger than the value under perfect foresight a). The ratio L_S/L depends on the long-run value of the rate of return r^*, the growth rate $n + \mu$, and the parameters ϵ and σ. When $r^* = 0.04$ and $n + \mu = 0.02$, it can be determined for different values of ϵ and σ, using table 2 and the relation (21).

We see in figure 3 that, for the relevant range of values of the various parameters, the ratio between the welfare gains of tax reform under myopic expectations and under perfect foresight is between .55 and .65 and is not very sensitive to the parameters of the model.[25]

Figure 3 also allows us to consider a more general class of expectations. For example, when the private sector underestimates the future rate of convergence of the economy, the realized rate of convergence a_I is in the interval (a, a_S). For this special class of expectations, the welfare gain of a tax cut is bounded below by the value found in the myopic case.

Finally, it may be interesting to observe that when the private sector applies a modest amount of rationality, the dynamic path of the economy is close to the optimal path. Around the steady state, the variation through time of the factor prices is given by the expression, $(z_t \dot{-} z^*) = -a_I(z_t - z^*)$.

Assume now that the long-run values of the factor prices are known and that the only uncertainty left is about their rate of convergence to the new steady state. Furthermore, individual expectations are of the regressive form; the expected values of a parameter x, at some future date t, x_t^e, are determined by the following rule:

$$(x_t^e \dot{-} x^*) = -\alpha(x_t^e - x^*), \tag{22}$$
$$x_0^e = x_0.$$

Individuals revise at each instant the estimated value α of the future rate of convergence, using the observed value $\hat{a}$ (given by $\hat{a} = \dot{z}/(z^* -$

[25] When the difference λ between the rate of return r^* and the growth rate $n + \mu$ tends to zero, the ratio L_S/L tends to 1/2. Also, at the same time, the error implied by myopic expectations, which is measured by the ratio a_s/a, tends to infinity.

z), where z is an arbitrarily chosen endogenous variable) and an adaptive rule, for example, $\dot{\alpha} = \nu(\hat{a} - \alpha)$, where $\nu > 0$.

Under these assumptions, and for an initial value of the capital stock not too different from the steady-state value, both the estimated value α and the actual value $\hat{a}$ of the rate of convergence tend to the optimal value a.[26] The dynamic path is *tangent* to the perfect foresight path at the steady-state point. The value of the welfare gain of tax reform is close to its (maximum) perfect-foresight value.

IV. The Welfare Cost of the Corporation Tax

Most studies on the corporation tax assume that the corporate and the noncorporate sectors produce two different goods and that the ratio between their respective outputs depends on their relative prices.[27] For the sake of simplicity, we assume here that both sectors produce the same good and that the aggregate production function can be written in the form,[28] $y = g(k_1, k_2)$, where all quantities are divided by the total effective labor supply, and k_1 and k_2 represent the corporate and the noncorporate capital stock, respectively. The function $g(k_1, k_2)$ is assumed to be homogeneous in its arguments.

Initially the economy is in the steady state where only the corporate capital income is taxed at the rate θ. In each sector, the net rate of return is equal to the optimal stationary value ρ^*, and the values of the capital stocks, $\bar{k}_1$ and $\bar{k}_2$, are determined by

$$\rho^* = (1 - \theta) \frac{\partial g}{\partial k_1} (\bar{k}_1, \bar{k}_2) = \frac{\partial g}{\partial k_2} (\bar{k}_1, \bar{k}_2). \tag{23}$$

The elimination of the corporate tax at time zero has two effects. First, capital is reallocated between the two sectors to equalize their

[26] Numerical simulations have shown that for almost any initial value of the capital stock, the convergence to the perfect-foresight path occurs after only a few periods. This procedure may be used as an algorithm to determine the perfect-foresight solution.

[27] See Harberger (1962, 1976), Shoven and Whalley (1972), Boadway and Treddenick (1975), and Shoven (1976) for a static analysis. Friedländer and Vandendorpe (1978) extend Harberger's study to a dynamic framework. However, they do not address the problem of the incidence on the intertemporal welfare and do not consider a saving function derived from the optimization of an intertemporal utility function. Fullerton, Shoven, and Whalley (1979) are currently working on a more elaborate model.

[28] The usual assumption that the corporate and the noncorporate sectors produce two separate goods may also be a crude description of reality; see, e.g., Ebrill and Hartman 1977.

rates of return. The aggregate production function now takes the form,

$$y = f(k), \tag{24}$$

where

$$f(k) = \max_{k_1 + k_2 = k} g(k_1, k_2).$$

After time zero, the private sector is endowed with perfect foresight, and the economy moves on the dynamic path studied in Section I, which is characterized by the equations, $\dot{c} = (c/\sigma)[f'(k) - \rho^*]$, $\dot{k} = f(k) - (n + \mu)k - c$, and $k_0 = \bar{k} = \bar{k}_1 + \bar{k}_2$. As in the relation (9), the welfare gain of the tax cut is equal to $\Delta U = J(\bar{k}) - [u(\bar{c})/(r^* - n - \mu)]$. This difference can be decomposed into

$$\Delta U = \left[J(\bar{k}) - \frac{u(\hat{c})}{r^* - n - \mu}\right] + \left[\frac{u(\hat{c}) - u(\bar{c})}{r^* - n - \mu}\right], \tag{25}$$

where $\hat{c} = f(\bar{k}) - (n + \mu)\bar{k}$ is the stationary consumption available when the level of capital $\bar{k}$ is allocated efficiently between the corporate and the noncorporate sectors.

The first term of the RHS (right-hand side) of (25) has been analyzed in Section II and measures the intertemporal welfare cost of the corporate tax, originating in the reduction of capital accumulation. It is equivalent to a permanent reduction of the consumption level by a fraction $L_{C1}\theta^2$, where L_{C1} is defined by[29]

$$L_{C1} = \left(\frac{\bar{k}_1}{\bar{k}}\right)^2 \cdot L. \tag{26}$$

Since the tax revenues are equal to $\theta r^* k_1$, this relation implies that for small values of the tax rates, the intertemporal welfare cost of capital income taxation depends only on the total amount of tax revenues (e.g., a 10 percent uniform tax on the whole capital stock and a 20 percent tax on half the capital stock have up to the second order the same intertemporal welfare cost).[30]

Although the distribution of the tax burden across the different sectors does not affect the intertemporal welfare cost, it creates a distortion in the allocation of the capital stock between sectors. This

[29] A proof is given in the Appendix, Sec. IV.

[30] The extension to many sectors is straightforward. In particular, this analysis could be applied to a model with four types of capital: corporate and noncorporate capital, housing, and human capital.

production cost is measured by the second term in the RHS of (25) and is equal to[31]

$$\frac{u(\hat{c}) - u(\bar{c})}{(r^* - n - \mu)} = \frac{u'(\bar{c})}{(r^* - n - \mu)} [f(\bar{k}) - g(\bar{k}_1, \bar{k}_2)]. \tag{27}$$

The welfare cost induced by the production inefficiency is equivalent to a permanent reduction of the consumption level by a fraction $L_{C2}\theta^2$, where

$$L_{C2}\theta^2 = [f(\bar{k}) - g(\bar{k}_1, \bar{k}_2)]/\bar{c}. \tag{28}$$

This is precisely the excess burden studied by Harberger and others.

The total welfare cost of the corporate tax is equal to the sum of the intertemporal cost and of the cost due to production inefficiency. It is equivalent to a permanent reduction of the consumption level by a fraction $L_C\theta^2$, where

$$L_C = L_{C1} + L_{C2} = L\left(\frac{\bar{k}_1}{\bar{k}}\right)^2 + \frac{[f(\bar{k}) - g(\bar{k}_1, \bar{k}_2)]}{\bar{c}}; \tag{29}$$

$\bar{k}_1$, $\bar{k}_2$, and $\bar{k}$ are given by (21), and

$$\bar{c} = f(\bar{k}) - (n + \mu)\bar{k}. \tag{30}$$

In order to have an order of magnitude of the respective quantities, assume that the ratio k_1/k is equal to $\frac{1}{3}$, which is close to the value observed in the U.S. economy, and that only the capital income originating in the corporate sector is taxed at the rate of 50 percent. Using the results of Section II, the intertemporal welfare cost is about 0.25 percent of the level of aggregate consumption (or around 7.8 percent of the tax revenues).[32] The value is somewhat lower than the cost of intersectoral misallocation, which in the same model is equal to 1.1 percent of consumption (this value is well within the range of current estimates of 0.5–1.5 percent; see Shoven 1976). The interindustry misallocation caused by the corporate tax seems to be much greater than the intertemporal distortion.

V. The Case of an Elastic Labor Supply

The capital income tax increases the gross rate of return and, by the factor price frontier, lowers the wage rate. When the labor supply is

[31] To a second order, it is equivalent to evaluate this difference at $\bar{k}$ or k^*.

[32] These estimates are obtained with a simplified version of the model in Section II: There is no depreciation, and the production function takes the form $y = k_1^{\beta}k_2^{\beta}$ (with a 50 percent corporate tax rate, $k_1/k_2 = 0.5$). The capital income share, net of depreciation, 2β, is the same as for the previous numerical results and is equal to 0.18. We have also used table 3 in the correction for large tax rates.

not fixed, this tax creates, in addition to the intertemporal distortion, a distortion in the choice between consumption and leisure at a given instant of time.

For example, assume that the elasticity of substitution between capital and labor in the production function is equal to zero; the capital-labor ratio is constant. When the labor supply is fixed, the tax has no incidence on the capital stock and is equivalent to a lump-sum tax with no excess burden. However, in the long run the gross rate of return is higher and, by the price possibility frontier, the wage rate is lower. When the labor supply depends on the net factor prices, the incidence of the tax is to decrease the long-run labor supply, the capital stock, and the consumption level.[33]

In general, it can be expected that the excess burden of the capital income tax is increased when the labor supply is variable. In this section, the framework developed in Section II is extended in order to measure this additional effect.

The utility function of the private sector depends on the amounts of consumption and leisure:

$$U = \int_0^\infty e^{-\rho t} e^{nt} u(c_t e^{\mu t}, l_t)dt, \tag{31}$$

where $c_t e^{\mu t}$ is the consumption per capita and l_t is the labor supply per capita (measured in natural units, and not in effective units).

For the sake of simplicity, we consider the following form for the utility function u:[34] $u(c,l) = (1 - \beta)\log c + \beta \log (T - l)$, where β and T are exogenous parameters. The dynamic path of the economy is now characterized by the relations,

$$\dot{c}_t = c_t[r_t(1 - \theta) - \rho^*], \tag{32}$$

$$\dot{k}_t = f(k_t, l_t) - (n + \mu)k_t - c_t, \tag{33}$$

and

$$\frac{\beta c_t}{(1 - \beta)(T - l_t)} = w_t, \tag{34}$$

where r_t and w_t are the gross interest and wage rates, respectively, and k_t represents now the aggregate capital stock divided by the efficiency index $e^{\mu t}$.

Equations (32) and (33) are the same as in the fixed labor supply case, and, at each time t, the marginal rate of substitution between the

[33] A sufficient condition for these properties to be verified is the homotheticity of the utility function. This homotheticity is a necessary condition for the existence of an optimal balanced growth path.

[34] When μ is different from 0, and the elasticity of substitution in u between consumption and leisure is different from 1, there is no optimal balanced growth path.

consumption of leisure and of produced goods is equal to the wage rate (eq. [34]).

At a given instant, the labor supply depends not only on the wage rate at the same moment but also on the future factor prices (wage and interest rates).

The initial state of the economy (where the tax has been in effect for an infinite amount of time) is defined by the stationary equivalents of (32)–(34):

$$\bar{r} = \frac{\rho^*}{1 - \theta}, \tag{35}$$

$$\bar{c} = A(\bar{r})\bar{l}, \tag{36}$$

and

$$\frac{\beta \bar{c}}{(1 - \beta)(T - \bar{l})} = w(\bar{r}), \tag{37}$$

where $A(r)$ is defined using the production function, $A(r) = r[(k/l)(r)] + w(r) - n - \mu$, and the capital-labor ratio k/l is expressed as a function of the gross rate of return r.

At time zero, the tax is suppressed. Thereafter, the dynamic behavior of the economy is described by (32), (33), and (34), where θ is replaced by 0; the new steady state is characterized by

$$r^* = \rho^*, \tag{38}$$

$$c^* = A(r^*)l^*, \tag{39}$$

and

$$\frac{\beta c^*}{(1 - \beta)(T - l^*)} = w(r^*). \tag{40}$$

The increase of welfare induced by the tax cut can be approximated to the second order by the same method as in Section II; its wealth equivalent is of the form, $\Delta M = L_\beta\, [c^*/(r^* - n - \mu)]\theta^2$, which depends on the parameter β. The formula for L_β is rather complicated,[35] so we present in tables 4 and 5 some numerical estimates analogous to those in Section II.

Also, L_β is represented as a function of ϵ in figure 2, for the values $\beta = 0, .6, 1$; β measures the long-run compensated elasticity of the labor supply with respect to the wage rate. When it is equal to zero, the labor supply is fixed; this is the case studied in Section II.

In table 4 and figure 2, we can see that the excess burden $L_\beta\theta^2$, for a given value of θ, is increasing with β and is bounded by the value of

[35] A computer program for its numerical evaluation is available.

TABLE 4

NORMALIZED WELFARE COST OF CAPITAL INCOME TAX WITH A VARIABLE LABOR SUPPLY (%)

	ϵ										
β	.0	.2	.4	.6	.8	1.0	1.2	1.4	1.6	1.8	2.0
.0	.0	1.21	2.33	3.41	4.45	5.46	6.44	7.40	8.33	9.25	10.14
.2	.45	1.60	2.71	3.77	4.81	5.82	6.80	7.76	8.71	9.63	10.54
.4	.86	1.98	3.07	4.13	5.16	6.17	7.16	8.13	9.09	10.03	10.96
.6	1.25	2.34	3.42	4.47	5.51	6.53	7.53	8.52	9.50	10.47	11.42
.8	1.61	2.68	3.75	4.81	5.85	6.89	7.92	8.93	9.94	10.95	11.94
1.0	1.95	3.01	4.07	5.13	6.20	7.26	8.32	9.39	10.45	11.51	12.58

TABLE 5

ANNUAL RATE OF CONVERGENCE TOWARD BALANCED GROWTH PATH WITH A VARIABLE LABOR SUPPLY (%)

	ϵ										
β	.0	.2	.4	.6	.8	1.0	1.2	1.4	1.6	1.8	2.0
.0	∞	23.74	16.51	13.31	11.40	10.10	9.14	8.40	7.80	7.30	6.88
.2	40.09	22.17	16.85	14.06	12.27	11.00	10.04	9.28	8.66	8.14	7.70
.4	28.47	20.91	17.20	14.90	13.31	12.11	11.18	10.42	9.79	9.25	8.79
.6	23.35	19.87	17.55	15.86	14.57	13.53	12.68	11.96	11.35	10.82	10.35
.8	20.32	19.01	17.91	16.97	16.16	15.45	14.83	14.27	13.76	13.30	12.89
1.0	18.28	18.28	18.27	18.27	18.26	18.26	18.25	18.24	18.24	18.23	18.23

$L_1\theta^2$, obtained for $\beta = 1$.[36] It is remarkable that the difference $L_\beta - L_0$ does not appear to be significantly affected by the value of ϵ[37] (for ϵ smaller than 2) and that this difference is relatively small with respect to L_0 when the production function is of the Cobb-Douglas type ($\epsilon = 1$).

When ϵ is infinite, we have the partial equilibrium case again. It is straightforward to show that the relation (14) is still valid.

The welfare cost of the capital income tax at the rate θ is measured by a fraction of the full consumption of produced goods and of leisure (valued at the wage rate), which is equal to $L_P\theta^2$, where $L_p = \frac{1}{2}[r^*/(r^* - n - \mu)]^2$.

VI. Conclusion

The general equilibrium models used in this study are highly stylized. However, the numerical examples considered indicate that the character of its results is fairly general.

The value of the excess burden of the capital income tax depends mainly on the elasticity of substitution between capital and labor ϵ. When ϵ is smaller than two, the excess burden is almost proportional to ϵ. No general rule could be derived about the effects of the discount rate. When the other parameters of the model are chosen to characterize the U.S. economy, the excess burden of the capital income tax increases with the discount rate and the growth rate.

Interestingly enough, when ϵ is smaller than two, an increase of the intertemporal substitutability of consumption (measured here by the inverse of the marginal utility of consumption) increases the value of the excess burden only by a negligible amount. In this case, although the deadweight loss of the tax is relatively small with respect to the level of aggregate consumption, or to its value in the case of fixed factor prices ($\epsilon = \infty$), it is not negligible with respect to the amount of tax revenues: A capital income tax of 50 percent implies a welfare loss equivalent to 26 percent of the tax revenues; when the tax rate is equal to 30 percent, the deadweight loss is equal to 11 percent of the tax revenues. These values should be multiplied by a number between 1 and 4/3 if the labor supply is elastic.

These numbers are indicative of the welfare gain induced by a shift from capital income taxation to lump-sum taxation, when individuals

[36] The compensated elasticity of the labor supply β is bounded by 1.

[37] Numerical experiments have shown that the difference $L_1 - L_0$ is not sensitive to variations of δ, ρ, or μ, for realistic values of ϵ. When ϵ is equal to zero, the excess burden of the capital income tax arises because of its incidence on the wage rate. In this special case, the interest tax has the same excess burden as the wage tax which generates an equal amount of revenues (Chamley 1980*a*).

have perfect foresight about the behavior of the economy after the tax is suppressed. Since the determination of the perfect-foresight path is not an easy task for most economists, even in a simple model, this assumption may not be realistic. When the tax reform does not lead to an optimal path, its benefits are reduced. For example, under myopic expectations, the welfare gain induced by the abolition of the capital income tax is only equal to about 60 percent of its value under perfect foresight. However, myopic expectations grossly underestimate the future behavior of the economy. When individuals anticipate somewhat this future behavior, the welfare gain of tax reform is close to its maximum value, even if the degree of foresight is not very accurate.[38] (A method to approximate the perfect foresight path has also been suggested in Sec. III.)

Finally, under the present U.S. tax system, the rates of the capital income tax vary by a large amount from one sector of production to another. When the overall elasticity of substitution between capital and labor is not too large (less than two), the welfare gains obtained by an equalization of these rates dwarf the gains obtained by a reduction of the global tax on capital income.[39]

Appendix

I. *The Excess Burden in Partial Equilibrium*

An income equivalent of the excess burden is given by the well-known Harberger-Hicks-Hotelling formula. In a continuous-time formulation, this becomes

$$\Delta M = \frac{1}{2}\int_0^\infty \int_0^\infty \Delta p_t \Delta p_{t'} \left(\frac{\partial c_t}{\partial p_{t'}}\right)^U e^{mt} dt dt', \tag{A1}$$

with the following notation:

$$\begin{cases} m = n + \mu \\ p_t = e^{-r^*(1-\theta)t} \text{ (price of } c_t \text{ at time 0)}, \\ \Delta p_t = r^*\theta t e^{-r^*t} \end{cases}$$

$\left(\frac{\partial c_t}{\partial p_{t'}}\right)^U$, compensated derivative of c_t with respect to $p_{t'}$.

[38] The assumption of myopic expectations is used in Hudson and Jorgenson (1976) and in Fullerton et al. (1979). Both studies rely on multisectoral models where one type of capital is allocated between the different sectors of production. Because there is only one state variable (the aggregate capital stock), regressive expectations are still asymptotically correct, and the discussion in Sec. III applies; in particular, the mechanism suggested there could be used to improve the accuracy of foresight.

[39] This result is also found in Fullerton et al. (1979).

By application of Slutsky's equation,

$$\Delta M = \frac{1}{2}\int_0^\infty e^{mt}\Delta p_t \left(\int_0^\infty \frac{\partial c_t}{\partial p_{t'}} \Delta p_{t'} dt'\right) dt$$

$$+ \frac{1}{2}\int_0^\infty e^{mt}\Delta p_t \frac{c_t}{M}\left(\frac{M}{c_t}\frac{\partial c_t}{\partial M}\right)\left(\int_0^\infty e^{mt'} c_{t'}\Delta p_{t'} dt'\right) dt.$$

The first expression in parentheses is simply equal to the variation of the consumption at time t, Δc_t, after an *uncompensated* interest tax has been applied (at time zero).

Since the utility function is homogeneous, the income elasticity of consumption is equal to one, and the excess burden is given by

$$\Delta M = \frac{1}{2}\int_0^\infty e^{mt}\Delta p_t \Delta c_t dt + \frac{1}{2M}\left(\int_0^\infty e^{mt} c_t \Delta p_t dt\right)^2 .$$

From (2) and the budget constraint, the uncompensated demand function, c_t, is expressed as follows: $c_t = \nu M e^{\alpha t}$, with $\nu = r^* - n - \mu - \theta r^*[1 - (1/\sigma)]$, and $\alpha = -(r^*/\sigma)\theta$. Therefore, when θ is small,

$$\Delta c_t = -Mr^* \left[\left(1 - \frac{1}{\sigma}\right) + \frac{(r^* - n - \mu)t}{\sigma}\right].$$

We replace Δc_t in the above expression for ΔM, and since the wealth M is equal to the present value of the consumption stream on the balanced growth path, we have

$$\Delta M = -L_P \frac{c^*}{r^* - n - \mu}\theta^2, \tag{A2}$$

with

$$L_P = \frac{1}{2\sigma}\left(\frac{r^*}{r^* - n - \mu}\right)^2.$$

For a given value of the interest tax rate, the welfare cost in partial equilibrium is an increasing function of the intertemporal substitutability of consumption in the utility function, which is measured by $1/\sigma$. It is a decreasing function of the discount rate and an increasing function of the growth rate.

II. *The Excess Burden in General Equilibrium*

We determine here a second-order equivalent of the welfare gain ΔU, obtained by the suppression of the interest tax: $\Delta U = J(\bar{k}) - [u(\bar{c})/\lambda]$ ($\lambda = \rho^* - n - \mu$). The terms $J(\bar{k})$ and $u(\bar{c})$ are considered separately.

The term $J(\bar{k})$ is equal to

$$J(\bar{k}) = \int_0^\infty e^{-\lambda t} u(c_t) dt,$$

where c_t is taken on the dynamic path BE (fig. 1). This integral can be decomposed into two terms:

$$J = u(c_t)\Delta t + \int_{\Delta t}^\infty e^{-\lambda t} u(c_t) dt,$$

where Δt is a small interval of time. By taking an infinitesimal value for Δt, we verify that the function J satisfies the relation, $0 = u(c) - \lambda J + J'(k)k$; or, using the capital accumulation equation (eq. [3]), $0 = u[c(k)] - \lambda J(k) + J'(k)[f(k) - (n + \mu)k - c(k)]$.

Differentiating this expression twice at the point k^*, we obtain

$$J'(k^*) = u'(c^*),\ J''(k^*) = \frac{u''c'^2 + f''u'}{\lambda + 2a},$$

where $u' = (du/dc)(c^*)$, $u'' = (d^2u/dc^2)(c^*)$, $c' = (dc/dk)(k^*)$, $c'' = [d^2c(k^*)]/dk^2$, and $a = c'(k^*) - \lambda$.

Therefore, a second-order approximation of J at $\bar{k}$ is given by

$$J(\bar{k}) = \frac{u}{\lambda} + u'(\bar{k} - k^*) + \frac{1}{2(\lambda + 2a)}(u''c'^2 + f''u')(\bar{k} - k^*)^2. \tag{A3}$$

Using the properties of the optimal consumption function described in Section I of the text, we can rearrange the term of the second order:

$$\frac{u''c'^2 + f''u'}{2(\lambda + 2a)} = \frac{u''}{2(\lambda + 2a)}\left(c'^2 - \frac{cf''}{\sigma}\right)$$

$$= \frac{u''(c'^2 + c'a)}{2(\lambda + 2a)} = \frac{u''c'(c' + \sigma a)}{2(\lambda + 2a)} = \frac{u''c'}{2}.$$

Hence,

$$J(\bar{k}) = \frac{u}{\lambda} + u'(\bar{k} - k^*) + \frac{u''c'}{2}(\bar{k} - k^*)^2. \tag{A4}$$

The value of consumption in the steady state under taxation, $\bar{c}$, is determined by $\bar{c} = f(\bar{k}) - (n + \mu)\bar{k}$. The second-order equivalent of $u(\bar{c})/\lambda$ can be written as

$$\frac{u(\bar{c})}{\lambda} = \frac{u}{\lambda} + u'(\bar{k} - k^*) + \frac{1}{2\lambda}(u'f'' + u''\lambda^2)(\bar{k} - k^*)^2, \tag{A5}$$

$$= \frac{u}{\lambda} + u'(\bar{k} - k^*) + \frac{u''}{2}\left(\lambda + \frac{c'a}{\lambda}\right)(\bar{k} - k^*)^2. \tag{A6}$$

Taking the difference between $J(\bar{k})$ and $u(\bar{c})/\lambda$ (in [A4] and [A6]), we have

$$\Delta U = -\frac{u''}{2}\frac{a^2}{\lambda}(\bar{k} - k^*)^2,$$

where a is the positive root of $x^2 + \lambda x - \gamma = 0$ (described in text Sec. I).

Since $f'(\bar{k}) = \rho^*/(1 - \theta)$ and $(\bar{k} - k^*)f'' \sim \rho^*\theta$, the welfare change can be rewritten as

$$\Delta U = \frac{u'}{2\sigma}\frac{c}{\lambda}\left(\frac{\rho^*}{\lambda}\right)^2\left(\frac{\sigma\lambda a}{cf''}\right)^2\theta^2 = u'\frac{c}{r^* - n - \mu}\theta^2 L,$$

where

$$L = L_P\left(\frac{\lambda a}{\gamma}\right)^2,$$

$$L_P = \frac{1}{2\sigma}\left(\frac{\rho^*}{\lambda}\right)^2$$

(the partial equilibrium value of L expressed by the relation [14] in the text), and

$$\gamma = \frac{-cf''(k^*)}{\sigma}.$$

The term $\lambda a/\gamma$ is equal to the positive root of the equation $(\gamma/\lambda^2)x^2 + x - 1 = 0$. This root is contained in the interval (0,1). The derivation of its properties when ϵ varies is straightforward. The sign of $\partial L/\partial\sigma$ is obtained by the same method.

III. *The Case of Nonperfect Foresight*

When the private sector is not endowed with perfect foresight, $c'(k^*)$ is no longer the positive root of the equation $x^2 - \lambda x - \gamma = 0$. The welfare gain of the interest tax cut is obtained by the same method as before. However, taking the difference between (A3) and (A5), we find

$$\Delta U = u'\left(\frac{\bar{k} - k^*}{2}\right)^2\left[f''\left(\frac{1}{\lambda + 2a_I} - \frac{1}{\lambda}\right) + \frac{\sigma}{c}\left(\lambda - \frac{c_I'^2}{\lambda + 2a_I}\right)\right],$$

where c_I and a_I represent the consumption function and the coefficient of adjustment of the economy toward the balanced growth path, under nonperfect foresight.

Use the relation $a_I = c_I' - \lambda$ and the definition of the perfect-foresight values for a_I and c_I', respectively a and c', to obtain

$$\Delta U = u'(c^*)f''(k^*)\left(\frac{\bar{k} - k^*}{2}\right)^2 Q(a_I),$$

where

$$Q(a_I) = \frac{a_I}{\lambda + 2a_I}\left(\frac{a_I}{ac'} - \frac{2}{\lambda}\right).$$

The consumption equivalent of the welfare gain can then be expressed as $L_I = L \cdot Q(a_I)/Q(a)$, where L is the value of L_I under perfect foresight.

IV. *The Corporation Tax*

We examine here the long-run incidence of the corporation tax on the capital stock. The capital stocks in the corporate and in the noncorporate sector, k_1 and k_2, respectively, are determined in the long run by the relations $\rho^* = (1 - \theta)g_1'(k_1, k_2) = g_2'(k_1, k_2)$. Differentiate this expression around $\theta = 0$ to obtain the variation of the aggregate capital stock, dk:

$$dk = dk_1 + dk_2 = \left(\frac{g_{22}'' - g_{12}''}{\Delta}\right)\rho^*\theta, \tag{A7}$$

with $\Delta = g_{11}''g_{22}'' - g_{12}''g_{21}''$. Assume now that the tax is applied uniformly to both sectors, and call Dk the variation of the capital stock:

$$Dk = \left(\frac{g_{11}'' + g_{22}'' - 2g_{12}''}{\Delta}\right)\rho^*\theta. \tag{A8}$$

Taking the ratio between (A7) and (A8),

$$\frac{dk}{Dk} = \frac{g_{22}'' - g_{12}''}{g_{11}'' + g_{22}'' - 2g_{12}''}.$$

Since the production function $g(k_1, k_2)$ is homogeneous, it can be rewritten under the form $g(k_1, k_2) = H[F(k_1, k_2)]$, where F is homogeneous of degree one.

Use the properties of F, and the equality of F'_1 and F'_2 (at $\theta = 0$), to write

$$g''_{22} - g''_{12} = H'(F''_{22} - F''_{12}) = H'F''_{22}\left(1 + \frac{k_2}{k_1}\right),$$

$$g''_{11} + g''_{22} - 2g''_{12} = H'F''_{22}\left[1 + 2\frac{k_2}{k_1} + \left(\frac{k_2}{k_1}\right)^2\right].$$

Taking the ratio between these relations, $dk/Dk = k_1/k$. This result could be generalized easily. The long-run incidence of a sectoral capital income tax on the total capital stock is proportional to the share of the taxed capital with respect to the total capital stock.

References

Arrow, Kenneth J., and Kurz, Mordecai. *Public Investment, the Rate of Return and Optimal Fiscal Policy.* Baltimore: Johns Hopkins Univ. Press, 1970.

Barro, Robert J. "Are Government Bonds Net Wealth?" *J.P.E.* 82, no. 6 (November/December 1974): 1095–1117.

Boadway, Robin, and Treddenick, John. "The Effects of the U.S. Corporate Tax on Resource Allocation and Welfare." Mimeographed. Kingston, Canada: Queen's Univ., March 1975.

Chamley, Christophe. "The Stability of the Capital Accumulation in a Competitive Economy with Imperfect Foresight." Mimeographed. New Haven, Conn.: Yale Univ., December 1979.

———. "Optimal Intertemporal Taxation and the Public Debt." Mimeographed. New Haven, Conn.: Yale Univ., April 1980. (*a*)

———. "Optimal Taxation in a Life-Cycle Capital Accumulation Model." Mimeographed. New Haven, Conn.: Yale Univ., May 1980. (*b*)

Diamond, Peter A. "National Debt in a Neoclassical Growth Model." *A.E.R.* 55 (December 1965): 1126–50.

———. "Incidence of an Interest Income Tax." *J. Econ. Theory* 2 (September 1970): 211–24.

Ebrill, Liam P., and Hartman, David G. "On the Incidence and Excess Burden of the Corporation Tax." Mimeographed. Cambridge, Mass.: Harvard Univ., October 1977.

Feldstein, Martin S. "Incidence of a Capital Income Tax in a Growing Economy with Variable Savings Rates." *Rev. Econ. Studies* 41 (October 1974): 505–13. (*a*)

———. "Tax Incidence in a Growing Economy with Variable Factor Supply." *Q.J.E.* 88 (November 1974): 551–73. (*b*)

———. "The Welfare Cost of Capital Income Taxation." *J.P.E.* 86, no. 2, suppl. (April 1978): S29–S51.

Friedländer, Ann F., and Vandendorpe, Adolf F. "Capital Taxation in a Dynamic General Equilibrium Setting." *J. Public Econ.* 10 (August 1978): 1–24.

Fullerton, Don; Shoven, John; and Whalley, John. "Dynamic General Equilibrium Impacts of Replacing the U.S. Income Tax with a Progressive Consumption Tax." Paper presented at the conference on the Taxation of Capital, Nat. Bur. Econ. Res., Cambridge, Mass., November 1979.

Green, Jerry R., and Sheshinski, Eytan. "Approximating the Efficiency Gain of Tax Reforms." *J. Public Econ.* 11 (April 1979): 179–195.

Hall, Robert E. "Consumption Taxes versus Income Taxes: Implications for Economic Growth." Paper presented at the 61st Annual Conference on Taxation, 1969.

———. "The Dynamic Effects of Fiscal Policy in an Economy with Foresight." *Rev. Econ. Studies* 38 (April 1971): 229–44.

Harberger, Arnold C. "The Incidence of the Corporation Income Tax." *J.P.E.* 70, no. 3 (June 1962): 215–40.

Harberger, Arnold C., and Bruce, Neil. "The Incidence and Efficiency Effects of Taxes on Income from Capital: A Reply." *J.P.E.* 84, no. 6 (December 1976): 1285–92.

Hudson, Edwards, and Jorgenson, Dale. "The Tax on Capital Income in the U.S." Report to Office of Taxation, U.S. Treasury Dept., Washington, January 1976.

Iwai, Katsuhito. "Optimal Economic Growth and Stationary Ordinal Utility—a Fisherian Approach." *J. Econ. Theory* 5 (August 1972): 121–51.

Jorgenson, Dale W., and Christensen, Laurits R. "Measuring Economic Performance in the Private Sector." In *The Measurement of Economic and Social Performance.* Studies in Income and Wealth, vol. 38. Edited by Milton Moss. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1973.

Koopmans, Tjalling C. "Objectives, Constraints and Outcomes in Optimal Growth Models." *Econometrica* 35 (January 1967): 1–15.

Krzyzaniak, Marian. "The Long-Run Burden of a General Tax on Profits in a Neoclassical World." *Public Finance* 22, no. 4 (1967): 472–91.

Laitner, John. "Household Bequests, Perfect Expectations, and the National Distribution of Wealth." *Econometrica* 47 (September 1979): 1175–93.

Levhari, David, and Sheshinski, Eytan. "Lifetime Excess Burden of a Tax." *J.P.E.* 80, no. 1 (January/February 1972): 139–47.

Sato, Kazuo. "On the Adjustment Time in Neo-classical Growth Models." *Rev. Econ. Studies* 33 (July 1966): 263–68.

———. "Taxation and Neo-classical Growth." *Public Finance* 22, no. 3 (1967): 346–70.

Shoven, John B. "The Incidence and Efficiency Effects of Taxes on Income from Capital." *J.P.E.* 84, no. 6 (December 1976): 1261–83.

Shoven, John B., and Whalley, John. "A General Equilibrium Calculation of the Effects of Differential Taxation of Income from Capital in the U.S." *J. Public Econ.* 1 (November 1972): 281–321.

Summers, Lawrence H. "Tax Policy in a Life-Cycle Model." Working Paper no. 302, Nat. Bur. Econ. Res., Cambridge, Mass., November 1979.

Weber, Warren E. "Interest Rates, Inflation, and Consumer Expenditures." *A.E.R.* 65 (December 1975): 843–58.

Wright, Colin. "Saving and the Rate of Interest." In *The Taxation of Income from Capital,* edited by Arnold C. Harberger and Martin J. Bailey. Washington: Brookings Inst., 1969.

Foreign Ownership and the Theory of Trade and Welfare

Richard A. Brecher

Carleton University

Jagdish N. Bhagwati

Columbia University

Some standard topics in the theory of international trade are reconsidered in this paper by distinguishing between national and aggregate income when fixed supplies of foreign inputs are present within the home country. Under conditions that would ensure a national welfare gain if foreign ownership were absent, international transfer, economic growth, or tariff policy might cause a national welfare loss in the presence of foreign ownership. The techniques developed could be applied to other domestic distinctions (such as those based on race, sex, age, or ethnicity) and to the theory of customs unions in a three-country world.

I. Introduction

This paper reconsiders a number of standard topics in the theory of international trade by taking explicit account of the distinction between national and aggregate income when fixed supplies of foreign-owned inputs are present within the domestic economy. Extending the work of Bhagwati and Brecher (1980),[1] the following

Thanks are due to the National Science Foundation grant no. SOC79-07541 for partial financial support of the research underlying this paper. The comments and suggestions of Alan Deardorff, Jacob Frenkel; Alasdair Smith, and anonymous referees are gratefully acknowledged.

[1] Their work, in turn, extends the analysis of Bhagwati and Tironi (1980), who concentrate on a special case mentioned in n. 3 below.

[*Journal of Political Economy*, 1981, vol. 89, no. 3]

analysis takes a new look at welfare-theoretic aspects of international transfer, economic expansion, and tariff policy, while it emphasizes significant departures from conventional wisdom that arise in the presence of foreign ownership. As these selected departures suggest, many standard results are open to serious question when part of the domestic product accrues to factor inputs from abroad.

Originally, the motivation for the present two-group analysis (based on the national-foreign distinction) came from a recent concern in Latin America, where policymakers have been worried about the impact of trade liberalization on national welfare, given the domestic presence of foreign-owned multinational corporations. After further reflection, however, it is clear that the treatment below has much greater applicability to a broad range of analytically similar cases. For example, it is possible to treat in much the same way a wide variety of alternative domestic distinctions, including those based on race, sex, age, or ethnicity. The following techniques and results, moreover, are directly relevant for the fully analogous two-group issue relating to the distribution of gains (or losses) between trading partners in a customs union (such as the European Economic Community) with factor mobility. While these other policy problems are of considerable importance and interest as well, only the national-foreign distinction is pursued explicitly here for the sake of brevity.

Section II reviews the basic model of an open economy, in which foreign-owned and national supplies of two homogeneous factors are combined to produce two commodities. As Section III then shows, a transfer-receiving country might suffer a loss in national welfare, even under the usual conditions which would ensure a welfare gain if foreign ownership were absent. As established next by Section IV, a country experiencing economic expansion (due to factor-endowment growth or technological advance) might encounter a deterioration in national welfare, even under well-known conditions which would preclude this possibility of "immiserizing growth" in the absence of foreign ownership. Afterward, Section V explains why free trade might be inferior to *both* no trade and subsidized trade, as far as national welfare is concerned.[2] Section VI summarizes the paper's main results, based on the possibility of aggregate and national welfare moving in opposite directions.

Needless to say, this possibility would not arise if foreign-owned factors were taxed to the nationally optimal extent. Indeed, with these factors in perfectly inelastic supply, the optimal tax on each foreign

[2] This result is obtained also by Bhagwati and Tironi (1980), for a special case identified in n. 3 below. In addition, since Bhagwati and Brecher (1980) compare free-trade equilibrium with autarky, the present paper will emphasize instead the comparison of free versus subsidized trade.

input clearly would be 100 percent, thereby removing the after-tax distinction between aggregate and national welfare. Assuming that this type of optimal taxation of factors is politically infeasible, however, the present analysis cautions nationally oriented policymakers against the usual, automatic adoption of the standard welfare conclusions which reflect an aggregate point of view. More specifically, this paper shows precisely how the traditional (aggregate) propositions must be modified for a truly national perspective, when political constraints eliminate optimal taxation of inputs from abroad.

II. The Basic Model

Following the analysis of Bhagwati and Brecher (1980), the present section summarizes the basic two-commodity, two-factor model of an open economy (large or small), which plays host to given quantities of inputs from abroad. The aggregate factor endowments of the country are fixed at $\bar{K}^a$ units of capital and $\bar{L}^a$ units of labor, while the given amounts $\bar{K}^n$ and $\bar{L}^n$ are the national endowments of capital and labor, respectively. (Thus, the fixed supplies of foreign-owned capital and labor within the home country are $\bar{K}^a - \bar{K}^n$ and $\bar{L}^a - \bar{L}^n$, respectively.) It is assumed that $\bar{K}^a > \bar{K}^n > 0$ and $\bar{L}^a > \bar{L}^n > 0$, excluding the possibility that either factor within the home economy is owned wholly by nationals or completely by foreigners.[3] Commodity two is always labor intensive relative to capital-intensive commodity one, and the well-behaved technology exhibits constant returns to scale.

In figure 1, the home country is depicted in free-trade equilibrium. Aggregate production is at point Q^a on production-possibility frontier $T_2^aT_1^a$ (corresponding to $\bar{K}^a$ and $\bar{L}^a$), aggregate income is represented by budget line Q^aD^a, and aggregate consumption occurs at point D^a on indifference curve $I_2^aI_1^a$. (For simplicity of exposition, it is assumed that all income earned by factors from abroad is consumed locally, to avoid having to show repatriation of such income within the diagram.) By the reasoning of Bhagwati and Brecher (1980), national consumption takes place at point D^n on indifference curve $I_2^nI_1^n$, with national income given by budget line Q^nD^n (parallel to Q^aD^a), *as if* nationals produced separately at point Q^n on production-possibility frontier $T_2^nT_1^n$ (drawn for $\bar{K}^n$ and $\bar{L}^n$).[4] To emphasize that the main results of this paper qualitatively do *not* require any differences in consumer preferences between nationals and foreigners within the home country, assume throughout the text that the same set of indifference

[3] For the special case in which $\bar{K}^a > \bar{K}^n = 0$ and $\bar{L}^a = \bar{L}^n > 0$, see Bhagwati and Tironi (1980).

[4] The discussion could be extended readily to allow for the possibility of complete specialization, following the analysis of Bhagwati and Brecher (1980).

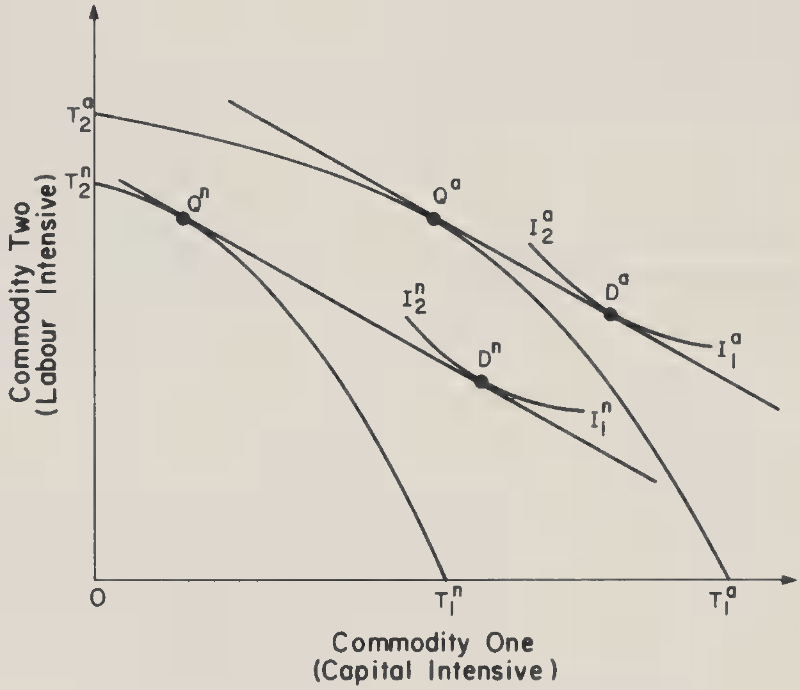


FIG. 1.—Differential trade-volume phenomenon

curves with unitary income elasticities of demand represents both national and aggregate tastes in consumption, although this simplification of the exposition could be dropped (as in footnotes to this paper) without detracting from the essence of the analysis.[5]

The model may be summarized conveniently as follows:

$$X^j_i = F^j_i(p), i = 1, 2, j = a, n; \tag{1}$$

$$Y^j = X^j_1 + pX^j_2,\ j = a, n; \tag{2}$$

$$W^j = U^j(C^j_1, C^j_2), j = a, n; \tag{3}$$

$$C^j_1 + pC^j_2 = Y^j, j = a, n; \tag{4}$$

where p denotes the relative price of the second commodity in terms of the first; X^j_i denotes output of commodity i on frontier $T^j_2T^j_1$; each F^j_i is a conventional function of p, given $\bar{K}^j, \bar{L}^j$ and the (uniform) technology for commodity i; Y^a and Y^n denote the real value of aggregate and national income, respectively, in terms of the first commodity; C^a_i and C^n_i denote aggregate and national consumption, respectively, of commodity i ($i = 1, 2$); W^a and W^n denote aggregate and national welfare,

[5] Nn. 6–8, 10, and 15 below extend the discussion to let tastes differ between nationals and foreigners within the home country. These extensions bring out the essentially "three-country" flavor of the analysis, in which nationals, domestically located foreigners, and the rest of the world can be treated as three distinct components of the international economy.

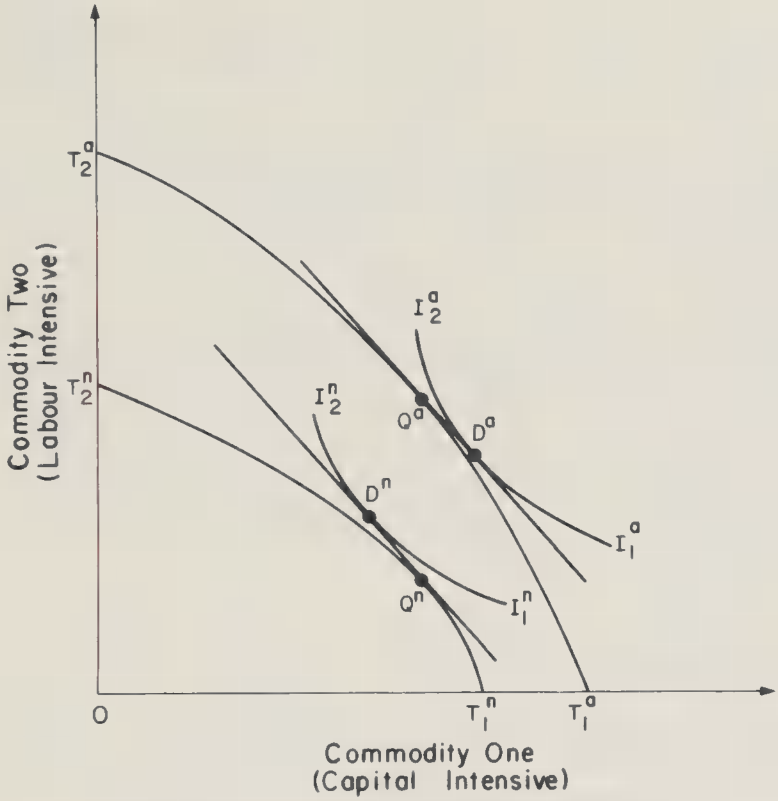


FIG. 2.—Differential trade-pattern phenomenon

respectively; and each U^j is a concave function of C_1^j and C_2^j, with positive partial derivatives denoted by $U_i^j \equiv \partial U^j / \partial C_i^j$ ($i = 1, 2$).

Later in the paper when a change in relative prices is induced by various parametric shifts, either of the following phenomena might lead to a fall in national despite a rise in aggregate welfare, depending on the strength of other induced effects. The differential trade-volume phenomenon is shown in figure 1, where the aggregate (actual) volume of trade (defined by line segment Q^aD^a) is less than the national (hypothetical) volume of trade (defined by line segment Q^nD^n), implying (ceteris paribus) that a terms-of-trade deterioration worsens national by more than aggregate welfare. Figure 2 (labeled similarly) illustrates the Bhagwati and Brecher (1980) differential trade-pattern phenomenon, which arises when the aggregate and national patterns of trade differ (in direction), so that an aggregate terms-of-trade improvement (tending to raise W^a) means a national terms-of-trade deterioration (tending to lower W^n). The national relative to the aggregate endowment of factors is labor abundant in figure 1 (with $\bar{K}^a/\bar{L}^a > \bar{K}^n/\bar{L}^n$) but capital abundant in figure 2 (with $\bar{K}^a/\bar{L}^a < \bar{K}^n/\bar{L}^n$), as suggested by the relative shapes of frontiers $T_2^aT_1^a$ and $T_2^nT_1^n$, in accordance with the reasoning of Rybczynski (1955).

To understand the possibility of a fall in national welfare despite a rise in aggregate welfare, it might be tempting to go no further than the following simple observation. Whenever the national and aggregate endowments exhibit different capital/labor ratios, the domestic distribution of income might deteriorate for nationals, as a change in relative commodity prices alters the wage/rental ratio for reasons expounded by Stolper and Samuelson (1941). It is important to recognize, however, that generally this income-redistribution effect will not be strong enough to produce the differential responses in national and aggregate welfare if the relative factor-endowment discrepancy is too small to create either the differential trade-volume or the differential trade-pattern phenomenon. Even when either of these phenomena arises, moreover, a fall in national welfare despite a rise in aggregate welfare can occur if and only if certain specific conditions (derived below) are satisfied.

III. International Transfer

According to a standard result in the literature (see Mundell 1960), a transfer-receiving country cannot suffer a loss in aggregate welfare despite any possible deterioration in the aggregate terms of trade, as long as international commodity-market equilibrium is stable. In other words, the transfer-induced change in W^a cannot be negative, assuming that an excess demand for or supply of the second good in world markets can be cleared by a rise or fall in p, respectively. As the following argument demonstrates, however, a (large) transfer-receiving country might suffer a deterioration in national welfare, even under the assumption (maintained throughout the present paper) that commodity markets are stable. This demonstration of a transfer-induced fall in W^n, moreover, does not even require a rise in the relative price of home importables.

If it is assumed that the transfer is given only to nationals, equations (2) are modified as follows:

$$Y^j = X_1^j + pX_2^j + \tau, j = a, n, \tag{5}$$

where τ is the real value of the transfer in terms of the first commodity. If any part of the transfer were given to foreigners within the home country, the chances for a decline in W^n would simply be enhanced, thereby strengthening the argument below.

To examine the welfare implications of the transfer, differentiate equations (1), (3), (4), and (5) totally with respect to τ—assuming (without loss of generality) that initially $U_1^j = 1$, while noting that $U_2^j/U_1^j = p = -(dF_1^j/dp)/(dF_2^j/dp)$ from the first-order conditions for

maximizing utility and profit. In this way, it is a straightforward exercise to derive

$$dW^j/d\tau = 1 + (E^j dp/d\tau), j = a, n, \tag{6}$$

where $E^j = X_2^j - C_2^j$. Consistent with figures 1 and 2, which depict the home country exporting the second good, $E^a > 0$ by assumption throughout the present paper. As illustrated above, however, E^n can be either positive (in fig. 1) or negative (in fig. 2).

As equations (6) confirm, $dW^a/d\tau$ is the familiar sum of the following two components: the primary gain (= 1) from the transfer-induced increase in aggregate income, at the initial (pretransfer) set of relative prices; plus the secondary effect ($= E^a dp/d\tau$) from the possible increase or decrease in the real exchange value of the initial volume of home exports, in the event of a transfer-induced change (if any) in relative prices. The expression for $dW^n/d\tau$ is analogous. If foreign inputs were entirely absent from the home country, the distinction between national and aggregate variables would disappear, thereby implying that $E^n = E^a$ and (hence) that $dW^n/d\tau = dW^a/d\tau$. Given the actual presence of factor inputs from abroad, however, $dW^n/d\tau$ generally differs from $dW^a/d\tau$, except in the special case where either $E^n = E^a$ (despite the foreign presence) or $dp/d\tau = 0$.

To determine precise conditions for the direction of change in welfare, consider the standard transfer-induced terms-of-trade response, analyzed previously by Samuelson (1952, 1954) and subsequently by Mundell (1960). Thus, by well-known reasoning,

$$dp/d\tau = (1 - m - m^*)/(e + e^* - 1)E^a, \tag{7}$$

where e (> 0) and m denote the relative-price elasticity of import demand and the marginal propensity to consume the importable, respectively, for the home country; e^* (> 0) and m^* denote the corresponding variables for the rest of the world; and $\tau = 0$ in the initial (pretransfer) equilibrium.[6] Given the assumption above that world commodity-market equilibrium is stable, $e + e^* > 1$ throughout the present paper.

If equation (7) is substituted into equations (6), simple manipulation confirms that

$$dW^a/d\tau = (\epsilon + \epsilon^*)/(e + e^* - 1) > 0, \tag{8}$$

[6] If preferences in consumption were allowed to differ between nationals and foreigners within the home country, it would be necessary to rewrite eq. (7) as follows, to reflect the present assumption that the entire transfer goes exclusively to nationals:

$$dp/d\tau = (1 - m^n - m^*)/(e + e^* - 1)E^a, \tag{7'}$$

where m^n denotes the national marginal propensity to consume the home importable.

but shows that

$$dW^n/d\tau \gtreqless 0 \text{ as } (e + e^* - 1)E^a \gtreqless (m + m^* - 1)E^n, \tag{9}$$

where ϵ (> 0) and ϵ^* (> 0) denote the compensated (constant-utility) relative-price elasticity of import demand for the home country and the rest of the world, respectively, while $e = \epsilon + m$ and $e^* = \epsilon^* + m^*$, according to a standard decomposition.[7] Although $dW^a/d\tau > 0$ unambiguously, it is evidently possible to have $dW^n/d\tau < 0$ nevertheless.[8]

To highlight the important role of the differential trade-pattern and differential trade-volume phenomena, it is helpful to revert to equations (6), which imply that a fall in national despite the rise in aggregate welfare can occur only if $(E^n - E^a)dp/d\tau < 0$. This necessary condition for a fall in W^n holds if either $dp/d\tau < 0$ in presence of the differential trade-volume phenomenon of figure 1 (where $E^n > E^a > 0$) or $dp/d\tau > 0$ in conjunction with the differential trade-pattern phenomenon of figure 2 (where $E^n < 0 < E^a$).[9] Correspondingly, if home exportables were relatively intensive in their use of capital (rather than labor), a transfer-induced deterioration in national

[7] Alternatively, if eq. (7′) from n. 6 above were substituted into eqq. (6), simple manipulation could show that

$$dW^a/d\tau = [(\epsilon + \epsilon^*) + (1 - \gamma)(m^f - m^n)]/(e + e^* - 1) \tag{8′}$$

and

$$dW^n/d\tau = [(\epsilon + \epsilon^*) + (1 - \gamma)(m^f + m^* - 1)]/(e + e^* - 1), \tag{9′}$$

where m^f denotes the marginal propensity to consume the home importable for foreigners within the home country; $\gamma \equiv (C_1^n - X_1^n)/(C_1^a - X_1^a) = E^n/E^a$; and use is made of the fact that $m = \gamma m^n + (1 - \gamma)m^f$. Eq. (8′) indicates that $dW^a/d\tau$ can be decomposed into two comparative-static components. As could be shown readily, the first component $[(\epsilon + \epsilon^*)/(e + e^* - 1)]$ is the transfer-induced change in W^a that would occur initially if the transfer were given temporarily to nationals and domestically located foreigners in the respective amounts $\gamma\tau$ and $(1 - \gamma)\tau$, whereas the second component $[(1 - \gamma)(m^f - m^n)/(e + e^* - 1)]$ is the subsequent change in W^a that would occur as the portion $(1 - \gamma)\tau$ was passed from domestically located foreigners to nationals (the ultimate recipients of the entire transfer). Eq. (9′) could be interpreted analogously, since $-dW^n/d\tau$ equals the worldwide sum of transfer-induced changes in welfare for everyone excluding home-country nationals, as could be shown readily.

[8] Under the present assumption that $m^f = m^n = m$, eq. (8′) of n. 7 above is equivalent to eq. (8), while eq. (9′) leads directly to condition (9). Alternatively, if it were the case that $m^f \neq m^n$, there would arise the new possibility of having $dW^a/d\tau < 0$ in eq. (8′). Also if it were supposed that $m^f = 1 - m^*$, it would be the case that $dW^n/d\tau > 0$ unambiguously in eq. (9′). This last result can be understood intuitively as follows: If foreign tastes are uniform throughout the world, the reasoning behind eq. (8) shows equally well that the transfer must lower worldwide foreign welfare; that is, $-dW^n/d\tau < 0$, recalling n. 7 above. Incidentally, in view of the fact that worldwide foreign welfare otherwise can rise (when $dW^n/d\tau < 0$) if $m^f \neq m^*$, international aid might be especially attractive for a donor country with investments in the aid-receiving economy.

[9] Although $(E^n - E^a)dp/d\tau < 0$ also if $0 < E^n < E^a$ when $dp/d\tau > 0$, $dW^n/d\tau > 0$ in this case, as implied by eqq. (6). The reader may also see alternatively that, from condition (9), $dW^n/d\tau < 0$ if and only if $(\epsilon + \epsilon^*)E^n + (E^a - E^n)(e + e^* - 1) < 0$. Therefore, national welfare *may* decline despite the increase in aggregate welfare if $E^n < 0 < E^a$ (i.e., the differential trade-pattern phenomenon holds) or if $E^n > E^a > 0$ (i.e., the differential trade-volume phenomenon holds).

(though not in aggregate) welfare would still be possible, provided that either the aggregate terms of trade improve in the case of labor-abundant nationals or an aggregate terms-of-trade decline occurs in the presence of capital-abundant nationals.

Consequently, the basic results of this section can be summarized generally in the following terms. When the home exportable uses intensively the factor that is relatively abundant in the national (as compared with the aggregate) endowment, the national and aggregate patterns of trade are the same, in which case a fall in national welfare might occur through a differential trade-volume phenomenon if the (national and aggregate) terms of trade worsen unambiguously. Alternatively, when the home exportable uses intensively the factor that is relatively scarce in the national (as compared with the aggregate) endowment, the aggregate and national patterns of trade could differ, in which case the differential trade-pattern phenomenon might give rise to a deterioration in national welfare if the national terms of trade worsen through an aggregate terms-of-trade improvement. These general results, moreover, hold equally well for changes in p induced by economic expansion and tariff policy, as will be clear from the analysis below.

IV. Economic Expansion

As Bhagwati (1958*a*) has demonstrated, a once-for-all increase in a factor endowment or in a technological level might deteriorate the aggregate terms of trade enough to worsen aggregate welfare of the home country, but this immiserizing growth can occur only if either the rest of the world has an inelastic offer curve or growth would decrease the production of home importables at the initial product-price ratio. In other words, if the offer-curve elasticity for the rest of the world is not less than unitary and economic expansion is not "ultrabiased" against the production of home importables, the growth-induced change in W^a cannot be negative. Even under these circumstances (assumed throughout the present section) which preclude a fall in aggregate welfare, however, the following analysis demonstrates that a (large) country might suffer a loss in national welfare. This demonstration of a growth-induced decline in W^n, moreover, does not even require a rise in the relative price of home importables.

To allow for factor-endowment expansion or technological advance, equations (1) may be rewritten as follows:

$$X^j_i = F^j_i(p, \theta), i = 1, 2, j = a, n, \tag{10}$$

where θ is a general shift parameter, a rise in which indicates either a factor-endowment increase (for $\bar{K}^j$ or $\bar{L}^j$) or a disembodied

technological improvement for an industry (one or two). It is assumed that any addition to the aggregate supply of capital or labor is owned fully by nationals. If any part of such addition were foreign owned, the likelihood of a decline in W^n would simply be enhanced, thereby strengthening the argument below. However, the ability of domestically located producers to take advantage of disembodied technological progress should be independent of the source of ownership of the inputs used, as assumed here.

Differentiating equations (2), (3), (4), and (10) totally with respect to θ, while recalling that $U_1^j = 1$ initially and that $U_2^j/U_1^j = p = -(\partial F_1^j/\partial p)/(\partial F_2^j/\partial p)$, we readily obtain the following result:

$$dW^j/d\theta = Y_\theta^j + (E^j dp/d\theta), j = a, n, \tag{11}$$

where $Y_\theta^j \equiv \partial Y^j/\partial\theta > 0$. Thus, each $dW^j/d\theta$ is the sum of a primary growth effect (Y_θ^j) plus a secondary relative-price effect ($E^j dp/d\theta$), which are analogous to the welfare-related effects of the transfer mentioned above in Section III. Although national and aggregate welfare again would remain equal if foreign inputs were entirely absent from the home country, the actual presence of foreign ownership gives rise to the possibility of having $dW^n/d\theta < 0$ when $dW^a/d\theta > 0$, except in the special case where $dp/d\theta = 0$.

Turning to the standard growth-induced terms-of-trade response, analyzed previously by Bhagwati (1958*b*) and subsequently by Kemp (1969, p. 110), we see that it is a well-known fact that

$$dp/d\theta = (\beta - m)Y_\theta^a/(e + e^* - 1)E^a, \tag{12}$$

where $\beta \equiv (\partial X_1^a/\partial\theta)/Y_\theta^a$. When this result is substituted into equations (11), straightforward manipulation confirms that

$$dW^a/d\theta = (\epsilon + \beta + e^* - 1)Y_\theta^a/(e + e^* - 1) > 0 \tag{13}$$

but shows that

$$dW^n/d\theta \gtreqless 0 \text{ as } E^a(e + e^* - 1)Y_\theta^n \gtreqless E^n(m - \beta)Y_\theta^a, \tag{14}$$

where $\beta \geqq 0$, which recalls the assumption that growth would not reduce production of home importables at the initial commodity-price ratio; and $e^* \geqq 1$, which recalls the assumption that the rest of the world's offer curve is not inelastic. Thus, despite the fact that $dW^a/d\theta > 0$ unambiguously under these circumstances, it is still possible to have $dW^n/d\theta < 0$ nevertheless.[10]

[10] Along lines suggested by nn. 6–8 for the case of international transfer, the analysis of economic expansion could be extended readily to distinguish between m^n and m^f. It is worth noting, however, that it would still be possible to have $dW^n/d\theta < 0$ even if it were the case that $m^f = 1 - m^*$.

Equations (11) imply that a fall in national despite the rise in aggregate welfare can occur only if $(E^n - E^a)dp/d\theta < Y^a_\theta - Y^n_\theta$. As could be shown readily, this necessary condition for a fall in W^n can result from each of the following alternative events, for example: an increase in either the national endowment of capital or the technological level of industry one, with $dp/d\theta > 0$ in the presence of the differential trade-pattern phenomenon (fig. 2); and an increase in either the national endowment of both factors or the technological level of both industries, if $dp/d\theta < 0$ with the differential trade-volume phenomenon (fig. 1).[11] Correspondingly, if home exportables were relatively capital intensive, it would be possible to have an expansion-induced deterioration in national (though not in aggregate) welfare under a variety of circumstances, including the following: an increase either in the national stock of labor or in the level of technology for the production of importables, when the national endowment is labor abundant; or an increase either in the national endowment of both factors or in the level of technology for both sectors, when nationals are capital abundant.

V. Tariff Policy

According to a standard result in the literature (see Bhagwati 1968), free trade is ranked superior to both no trade and subsidized trade (assuming that both offer curves are well behaved),[12] from the viewpoint of aggregate welfare. In other words, the home country cannot increase W^a above the free-trade level either by using an import (or export) tax to eliminate trade or by imposing an export (or import) subsidy to encourage trade. From the national-welfare point of view, however, the ranking above may be reversed. Since Bhagwati and Brecher (1980) already demonstrated the possibility of such a reversal for free trade versus autarky, the following analysis concentrates on free versus subsidized trade.

To allow for tariff policy, equations (2) may be modified as follows:

$$Y^j = X^j_1 + pX^j_2 + [(C^a_1 - X^a_1)\,\alpha/(1 - \alpha)], j = a, n; \tag{15}$$

where α denotes the *ad valorem* tariff, which is an import tax (if $\alpha > 0$) or an import subsidy (if $\alpha < 0$); the domestic relative price of the

[11] Although $Y^a_\theta = Y^n_\theta$ with a national factor-endowment increase, it is possible that $Y^a_\theta > Y^n_\theta$ for a technological advance. Thus, with the latter (but not the former) type of economic expansion, a fall in national welfare despite the rise in aggregate welfare might occur even without the differential trade-volume and differential trade-pattern phenomena—if both industries experience the technological advance and $dp/d\theta < 0$.

[12] I.e., the offer curve is assumed to represent imports as a monotonic decreasing function of their relative price in world markets. For the significance of this assumption in tariff analysis, see Bhagwati and Kemp (1969).

second good is still denoted by p, so that the relative price of this good in world markets is now equal to $p(1 + \alpha)$; and $(C_1^a - X_1^a)\alpha/(1 + \alpha)$ equals the real value (in terms of the first good) of tax revenues or subsidy payments, evaluated at domestic prices.[13] In writing equations (15), it is assumed (for the sake of simplicity) that all tax revenues or subsidy payments, respectively, are returned to or collected from *national* consumers in lump-sum fashion. If foreigners within the home country were to receive or finance any part of these revenues or payments, respectively, comparison of the free-trade and autarkic equilibria (which generate no tax revenues) clearly would be unaffected, while the chances of having free trade inferior to subsidized trade simply would be enhanced (thereby strengthening the analysis below).

To show that free trade might be inferior to subsidized trade from the national point of view, it is sufficient to establish the possibility of having $dW^n/d\alpha < 0$ in free-trade equilibrium. Consequently, throughout the following discussion, let $\alpha = 0$ in the initial (pretariff) equilibrium.

Differentiating equations (1), (3), (4), and (15) totally with respect to α and again recalling that $U_1^j = 1$ initially and that $U_2^j/U_1^j = p = -(dF_1^j/dp)/(dF_2^j/dp)$, we may verify readily that

$$dW^j/d\alpha = pE^a + (E^j dp/d\alpha), j = a, n; \tag{16}$$

note that $pE^a = C_1^a - X_1^a$ when (balanced) trade is initially free (with $\alpha = 0$). By well-known reasoning (see Kemp 1969, p. 96),

$$dp/d\alpha = p(1 - m - e^*)/(e + e^* - 1); \tag{17}$$

note that $1 - m$ equals the home country's marginal propensity to consume the exportable and recall that $\alpha = 0$ initially. When this result is substituted into equations (16), simple manipulation confirms that[14]

$$dW^a/d\alpha = \epsilon pE^a/(e + e^* - 1) \geqq 0 \tag{18}$$

but shows that

$$dW^n/d\alpha \gtreqless 0 \text{ as } (e + e^* - 1)E^a \gtreqless (m + e^* - 1)E^n. \tag{19}$$

Thus, despite the fact that $dW^a/d\alpha \geqq 0$, it is evidently possible to have $dW^n/d\alpha < 0$ nevertheless.[15]

[13] Although the corresponding value at *world* prices would be $(C_1^a - X_1^a)\alpha$, consumers respond directly to *domestic* (tariff-inclusive) prices instead.

[14] Note that $dW^a/d\alpha = 0$ only in the small-country case where $e^* = \infty$, and even then the change in W^a does not equal zero for any discrete change in α, by well-known reasoning.

[15] Along lines suggested by nn. 6–8, the analysis of tariff policy could be extended readily to let $m^f \neq m^n$, without eliminating the possibility of having $dW^n/d\alpha < 0$ even if $m^f = 1 - m^*$.

As implied by the equations (16), it is possible to have $dW^n/d\alpha < 0$ (even though $dW^a/d\alpha$ cannot be negative) if either a differential trade-volume phenomenon arises (fig. 1) when $dp/d\alpha < 0$ (the "normal" price response) or a differential trade-pattern phenomenon occurs (fig. 2) when $dp/d\alpha > 0$. (The "perverse" price response [$dp/d\alpha > 0$] can occur only in the large-country case, under conditions discussed by Metzler [1949].) Correspondingly, if home exportables were relatively capital intensive, it would be possible to have $dW^n/d\alpha < 0$ (even though $dW^a/d\alpha$ cannot be negative) if either nationals are labor abundant when $dp/d\alpha > 0$ or nationals are capital abundant when $dp/d\alpha < 0$.

Thus a trade subsidy might raise national (but not aggregate) welfare above the free-trade level. This analysis of a small subsidy (tax) on trade, moreover, complements the discussion of Bhagwati and Brecher (1980), who concentrate on prohibitive taxes on trade and thus are able to avoid the issue of tariff revenues.

The analysis of this section has an important implication for the traditional method of estimating the cost (benefit) of tariff protection or trade liberalization. Since the conventional method (as outlined by Johnson [1960]) ignores the source of ownership of domestically located inputs, the concept measured (in present notation) is clearly $dW^a/d\alpha$ rather than $dW^n/d\alpha$. Thus, the traditional estimate of the impact of protection or liberalization is an aggregate measure, which overstates or understates the national cost (benefit) if $(E^n - E^a)dp/d\alpha \lesseqgtr 0$, respectively, as suggested by equations (16). This misstatement arises because the conventional estimate simply sums the three standard components (namely, the external terms-of-trade effect and the costs of distortion in both production and consumption), while it fails to exclude the foreign-factor portion of the tariff-induced change in aggregate welfare.[16]

VI. Summary

As demonstrated by this paper, welfare aspects of international trade theory need to be reconsidered, when national and aggregate income

[16] The foreign-factor portion of the change in aggregate welfare is represented by the expression $(E^a - E^n)dp/d\alpha$, which must be excluded from $dW^a/d\alpha$ to give $dW^n/d\alpha$, as suggested by eqq. (16). Also by repeating the procedure of Bhagwati, Ramaswami, and Srinivasan (1969), it is possible to write (in present notation) that $dW^a/d\alpha = (E^a dp^*/d\alpha) + [(p^* - p)dX_2^a/d\alpha] + [(p - p^*)dC_2^a/d\alpha]$, where the relative price of the second good in world markets is denoted p^*, which equals $p(1 + \alpha)$. The components $E^a dp^*/d\alpha$, $(p^* - p)dX_2^a/d\alpha$, and $(p - p^*)dC_2^a/d\alpha$ are the effects due to the terms-of-trade change, the production distortion, and the consumption distortion, respectively. When evaluated in free-trade equilibrium (where $\alpha = 0$ and $p = p^*$), the latter two (distortion-related) components disappear, leaving only the first (terms-of-trade) component. This remaining (first) component, moreover, is equivalent to the right-hand side of eqq. (16) for $j = a$, since (in free-trade equilibrium) $dp/d\alpha = (dp^*/d\alpha) - p$.

differ in the presence of foreign ownership. Examples of this need are provided by the analysis of international transfer, economic expansion, and tariff policy. For a country receiving a transfer from abroad, national (but not aggregate) welfare might deteriorate even when international commodity-market equilibrium is stable, regardless of the direction of change in the world product-price ratio. In the case of economic expansion from factor-supply growth or technological advance, national (but not aggregate) welfare might worsen even when the rest of the world does not have an inelastic offer curve and domestic expansion is not ultrabiased against production of home importables, no matter what the direction of change in the world commodity-price ratio. As for tariff policy, free trade might be ranked inferior to both no trade and subsidized trade (in either direction), from the viewpoint of national (but not aggregate) welfare. Moreover, the conventional empirical estimates of the cost of protection à la Johnson's (1960) methodology are generally seen to be in need of correction if the economy has foreign-owned factors of production. In fact, many economies typically do have substantial labor inflows under *gastarbeiter* or other programs defined by immigration-quota policies, and, of course, equally there are substantial flows of capital among nation states.[17]

References

Bhagwati, Jagdish N. "Immiserizing Growth: A Geometrical Note." *Rev. Econ. Studies* 25 (June 1958): 201–5. (*a*)

———. "International Trade and Economic Expansion." *A.E.R.* 48 (December 1958): 941–53. (*b*)

———. "The Gains from Trade Once Again." *Oxford Econ. Papers* 20 (July 1968): 137–48.

———. "Shifting Comparative Advantage, Protectionist Demands, and Policy Response Options." Paper presented at a Ford-NBER conference on Import Competition and Adjustment: Theory and Policy, Cambridge, Mass., May 1980.

Bhagwati, Jagdish N., and Brecher, Richard A. "National Welfare in an Open Economy in the Presence of Foreign-owned Factors of Production." *J. Internat. Econ.* 10 (February 1980): 103–15.

Bhagwati, Jagdish N., and Kemp, Murray C. "Ranking of Tariffs under Monopoly Power in Trade." *Q.J.E.* 83 (May 1969): 330–35.

Bhagwati, Jagdish N.; Ramaswami, V. K.; and Srinivasan, T. N. "Domestic Distortions, Tariffs, and the Theory of Optimum Subsidy: Some Further Results." *J.P.E.* 77 (November/December 1969): 1005–10.

[17] Variations in protection may, in turn, lead to variations in the presence of foreign factors in the economy. Thus, for instance, Bhagwati (1980) has considered linkages between reduction in protection and reduction in the restrictiveness of immigration quotas. However, the analysis in the text has taken the endowment of *both* national and foreign-owned factors to be invariant to policy changes.

Bhagwati, Jagdish N., and Tironi, Ernesto. "Tariff Change, Foreign Capital and Immiserization: A Theoretical Analysis." *J. Development Econ.* 7 (March 1980): 71–83.

Johnson, Harry G. "The Cost of Protection and the Scientific Tariff." *J.P.E.* 68 (August 1960): 327–45.

Kemp, Murray C. *The Pure Theory of International Trade and Investment.* Englewood Cliffs, N.J.: Prentice-Hall, 1969.

Metzler, Lloyd A. "Tariffs, the Terms of Trade, and the Distribution of National Income." *J.P.E.* 57 (February 1949): 1–29.

Mundell, Robert A. "The Pure Theory of International Trade." *A.E.R.* 50 (March 1960): 67–110.

Rybczynski, T. M. "Factor Endowment and Relative Commodity Prices." *Economica* 22 (November 1955): 336–41.

Samuelson, Paul A. "The Transfer Problem and Transport Costs: The Terms of Trade When Impediments Are Absent." *Econ. J.* 62 (June 1952): 278–304.

———. "The Transfer Problem and Transport Costs, II: Analysis of Effects of Trade Impediments." *Econ. J.* 64 (June 1954): 264–89.

Stolper, Wolfgang F., and Samuelson, Paul A. "Protection and Real Wages." *Rev. Econ. Studies* 9 (November 1941): 58–73.

Accounting for Price Changes: American Steel Rails, 1879–1910

Robert C. Allen

University of British Columbia

A framework is developed for decomposing product price changes into changes in input prices, technical efficiency, and deviations of price from unit cost. This framework facilitates the measurement of productivity growth in noncompetitive industries. The history of American steel rail prices between 1879 and 1910 is analyzed, and it is concluded (in contrast with much recent work) that productivity growth remained rapid until the twentieth century and that the steel industry was sufficiently collusive so that the rail producers received the benefits of that productivity growth as excess profits.

I

The 1890s mark a pronounced change in the position of the American steel industry in the world market. Before the 1890s American prices substantially exceeded British prices, and the American industry achieved a large size only because of high tariffs. During the 1890s American prices dropped to British levels or below, and America emerged as a major exporter of iron and steel. This change in price is apparent in the case of rails, which was one of the leading steel products of the time. Between 1881 and 1890 the average price of steel rails at Pennsylvania mills was $37.01 while the average British price was $23.62. During the period 1906–13 the American price had fallen to $28.00 while the British price had risen to $29.46 (Allen

I am extremely grateful to W. E. Diewert for his freely given assistance during the gestation period of this paper. I also thank Michael B. Percy, Peter Chinloy, C. F. J. Boonekamp, Frank Lewis, Peter Lindert, E. R. Berndt, and David Donaldson for their comments on earlier versions of the paper.

[*Journal of Political Economy,* 1981, vol. 89, no. 3]

1979). To explain the emergence of the United States as a steel exporter, it is necessary to explain the fall in American steel prices.[1] It is the aim of this paper to account for the decline in the price of Bessemer steel rails.

There is no shortage of possible explanatory factors. During the late nineteenth century low-cost iron ore fields—particularly the Mesabi in Minnesota—were exploited, and improvements in lake shipping services were effected. Both developments substantially lowered the cost of iron ore in the Midwest and hence the cost of Bessemer pig iron, which was the principal input in steel rails. Moreover, many observers at the time believed the rate of productivity growth to have been very high in the American steel industry. This development would also be expected to have lowered the price of American steel. On the other hand, there were several attempts on the part of producers to collude and thus raise the price of steel rails. With the formation of U.S. Steel in 1901 the difficulty of enforcing collusive agreements was reduced. In the case of steel rails there is prima facie evidence that producers subsequently exercised monopoly power. Prior to 1901, the price of steel rails had changed frequently. In that year, however, the price of steel rails was set at $28 per ton, and it remained unchanged until the First World War. Such stability was unprecedented in the industry and indicates some power to control the price but does not in itself prove that price was set above cost (Temin 1964, pp. 192, 284–85).

The procedures that economic historians have used to analyze price changes have not been fully satisfactory. The technique most firmly rooted in theory is the application of an identity derived by Jorgenson (1966, pp. 2–4). The identity asserts that the rate of change of the price of the product equals an index of the rates of change of input prices minus a residual which is usually interpreted as the rate of technical progress. McCloskey (1973, p. 99) used a relative levels version of this identity to argue that—since the ratio of the price of rails to the price of pig iron (which was regarded as a proxy for the index of input prices) was trendless after the 1880s—there was no productivity growth after that date.[2] The difficulty with this procedure is that the identification of the residual with technical progress is correct only if the industry is in competitive long-run equilibrium. This assumption was barely mentioned in Jorgenson's discussion of the identity,[3] and, indeed, it is not immediately obvious from the

[1] It is also necessary to explain the rise in British prices. This aspect of the problem is considered in Allen (1979).

[2] Temin (1964, pp. 218–19) develops a similar argument.

[3] The discussion is confined to Jorgenson (1966, p. 4, n. 1).

mathematics that the assumption has been invoked. However, a scheme which decomposes price changes into efficiency and input price effects only if the industry is in competitive long-run equilibrium is not a satisfactory framework for analyzing American steel prices in the late nineteenth and early twentieth centuries.

Reflection suggests that changes in market power influence the difference between input and output prices. This realization has also been exploited by economic historians analyzing the evolution of steel prices. Thus McCloskey (1973, pp. 24–28) examined the ratio of rail to pig iron prices in order to investigate the effectiveness of various attempts at collusion by rail producers in Britain. Temin (1964, pp. 186–92) undertook similar investigations for the United States. Clearly, changes in monopoly power and technical progress cannot both be inferred from the ratio of output to input prices. At best, only their combined effect can be identified.

What is required to account for the fall in the price of American steel is an accounting scheme that identifies the separate effects of input price changes, efficiency changes, and changes in the deviation of price from unit cost. Such an accounting scheme is developed in Section II for a single product firm (an analogous scheme for a multiproduct firm is developed in an appendix).[4] In Section III the scheme is used to account for the fall in the price of steel rails between 1879 and 1910. The data discussed there indicate that declines in input prices and increases in efficiency both exerted powerful downward pressures. Indeed, the rate of technical progress was at its highest in the 1890s. The greater effectiveness of collusion after the U.S. Steel merger, however, meant that these cost declines were not entirely reflected in lower prices. Indeed, profit margins were raised enough between 1889 and 1902 so that rail producers realized most all of the gains of technical progress as higher profits. In Section IV these conclusions are compared with McCloskey's and Temin's findings. Jorgenson's price accounting identity is examined in detail in order to make explicit the necessity of assuming competitive long-run equilibrium if one tries to infer the rate of productivity change from differences in the rates of change of input and product prices. It is also shown that substantial errors result if this assumption is erroneously maintained.

[4] The author has extended the price accounting framework discussed here to the case of a multiple output firm. The multiple output case is developed in Allen (1978, pp. 7–11) and in an appendix to this paper which is available from the author upon request.

II

To develop the price accounting scheme for a single product firm, it is assumed that the firm minimizes costs subject to the following production function:

$$Q(t) = A(t) \cdot f[X(t)], \tag{1}$$

in which $Q(t)$ is output at time t, $X(t) = [X_1(t), X_2(t), \ldots, X_k(t)]$ is the vector of inputs used at time t, and $A(t)$ is an index of Hicks neutral technical progress. The assumption of Hicks neutrality can easily be relaxed (Diewert 1980, pp. 489–95). The function f is assumed to be linearly homogeneous. If f were known, the relative increase in efficiency between t_2 and t_1 could be computed from $Q(t_1)$, $Q(t_2)$, $X(t_1)$, and $X(t_2)$:

$$\frac{A(t_2)}{A(t_1)} = \frac{Q(t_2)/Q(t_1)}{f[X(t_2)]/f[X(t_2)]}. \tag{2}$$

The term $Q(t_2)/Q(t_1)$ is the relative increase in output that actually occurred. The term $f[X(t_2)]/f[X(t_1)] = A(t_1) \cdot f[X(t_2)]/A(t_1) \cdot f[X(t_1)]$ is the relative increase in output that would have occurred in the absence of technical change but given the change in input quantities that actually occurred. Thus $A(t_2)/A(t_1)$ equals the proportional increase in output that is not attributable to the increase in inputs.

If the firm minimizes cost, then there exists a total cost function C which is factorable into a unit cost function c^* multiplied by the production rate. Under the assumption that technical progress is Hicks neutral, the unit cost function can be written as a function of the input prices divided by $A(t)$:

$$C[t,W(t),Q(t)] = c^*[t,W(t)] \cdot Q(t) = c[W(t)] \cdot Q(t)/A(t). \tag{3}$$

Here $W(t) = [W_1(t), W_2(t), \ldots, W_k(t)]$ is the vector of k input prices. It is presumed that the supply to the firm of $X(t)$ is perfectly elastic at $W(t)$. The right-hand equality implies that the relative change in unit costs can be decomposed into an efficiency component and an input price component:

$$\frac{c^*[t_2,W(t_2)]}{c^*[t_1,W(t_1)]} = \frac{A(t_1)}{A(t_2)} \cdot \frac{c[W(t_2)]}{c[W(t_1)]}. \tag{4}$$

The term $A(t_2)/A(t_1)$ is relative efficiency in the two periods and $c[W(t_2)]/c[W(t_1)]$ is the relative increase in unit cost that would have obtained had there been no productivity growth but given the actual change in input prices. Equation (4) provides a second method for measuring productivity growth. If c were known, then $c[W(t_2)]/c[W(t_1)]$

could be computed by direct substitution. Dividing that result by the observed change in unit costs would yield $A(t_2)/A(t_1)$. From the perspective of equation (4), the increase in efficiency equals the relative decline in unit costs not attributable to declining input prices.

The price of the product will always be related to unit costs by the following identity:

$$P(t) = M(t) \cdot c^*[t, W(t)], \tag{5}$$

where $M(t)$ is the implicitly defined markup of price over cost. For a competitive industry in long-run equilibrium, $M(t) = 1$. One would expect to observe $M(t)$ usually not equal to one. The relative change in price between t_1 and t_2 can be written as

$$\frac{P(t_2)}{P(t_1)} = \frac{M(t_2)}{M(t_1)} \cdot \frac{c^*[t_2, W(t_2)]}{c^*[t_1, W(t_1)]}. \tag{6}$$

Equation (6) could be used to decompose a change in price into a cost change and a markup change. Equations (2) and (4) provide a basis for decomposing the cost change into input price and efficiency components.

If c and f were known, equations (2), (4), and (6) could be used to decompose price changes into markup, efficiency, and input price changes. In an empirical study, of course, the functional forms of c and f would have to be specified and the parameters of the function estimated. Prodigious task! Recent work in index number theory (e.g., Diewert 1976, 1980) has shown how such elaborate undertakings can be circumvented. Corresponding to each of the functional forms that would likely be used to specify c or f are one or more index number formulae. The indexed input quantities (or input prices) would indicate exactly $f[X(t_2)]/f[X(t_1)]$ or $c[W(t_2)]/c[W(t_1)]$ on the assumption that f or c had the functional form that corresponded to the index number formula used. For example, $f[X(t_2)]/f[X(t_1)]$ would equal a Laspeyres or Paasche index of the input quantities if f were Leontief. If f were Cobb-Douglas, $f[X(t_2)]/f[X(t_1)]$ would equal a geometric index of input quantities. If f were a quadratic function, $f[X(t_2)]/f[X(t_1)]$ would equal a Fisher ideal index, and if f were translog, then $f[X(t_2)]/f[X(t_1)]$ would equal a Törnqvist index of the input quantities.[5] Analogous results

[5] These equivalences are discussed in Diewert (1980, pp. 446–52). The linearly homogeneous translog function, which is used extensively in this paper, is

$$\ln f[X(t)] = \alpha_0 + \sum_{i=1}^{k} \beta_i \ln X_i(t) + 1/2 \sum_{m=1}^{k} \sum_{n=1}^{k} \beta_{mm} \ln X_m(t) \ln X_n(t),$$

subject to the restrictions: $\Sigma_{i=1}^{k} \beta_i = 1$, $\beta_{mn} = \beta_{nm}$, and $\Sigma_{i=1}^{k} \beta_{mn} = 0$ for $m = 1, 2, \ldots, k$. Diewert (1976) has shown that

hold for c and input price indices. It should be noted that in the index number approach to the measurement of $f[X(t_2)]/f[X(t_1)]$ or $c[W(t_2)]/c[W(t_1)]$ it is not necessary to know the values of the parameters of f or c. Instead, the index number formulae require the values of the factor shares, estimates of which are frequently easier to obtain than estimates of the parameters of f or c.

In an empirical study, one should choose the formula for indexing input prices or quantities that corresponds to the true structure of c and f. In the absence of any prior information on the true structure, the choice of an index number formula must be made on the basis of which possible specification of f or c would best approximate the true structure, whatever it is. Since Leontief and Cobb-Douglas functions can be regarded as first-order approximations to f or c, while quadratic and translog functions can be regarded as second-order approximations, one would expect the quadratic and translog functions to fit the data more closely. Hence, there are grounds for preferring the Törnqvist and the Fisher ideal indices to the Laspeyres, Paasche, or geometric. Since the Fisher ideal and Törnqvist indices are exact for functions that can be regarded as second-order approximations to any linearly homogeneous production, cost, or expenditure function, Diewert (1976) has termed the Törnqvist and Fisher ideal indices "superlative." In the remainder of this paper, translog functions and Törnqvist indices will be used.

There are two ways to proceed, depending on whether one measures $A(t_2)/A(t_1)$ with equation (2) or (4). If one uses equation (2), one presumes f is translog, so $f[X(t_2)]/f[X(t_1)]$ equals a Törnqvist index of inputs. Hence,

$$\frac{A(t_2)}{A(t_1)} = \frac{Q(t_2)/Q(t_1)}{\prod_{i=1}^{k}\left[\frac{X_i(t_2)}{X_i(t_1)}\right]^{\bar{s}_i}} = \prod_{i=1}^{k}\left[\frac{Q(t_2)/X_i(t_2)}{Q(t_1)/X_i(t_1)}\right]^{\bar{s}_i}. \tag{7}$$

Consistency requires that $c[W(t_2)]/c[W(t_1)]$ be measured implicitly by substituting $A(t_2)/A(t_1)$ from equation (7) into equation (5):

$$\tilde{c}[W(t_2),W(t_1)] = \frac{c^*[t_2,W(t_2)]}{c^*[t_1,W(t_1)]} \cdot \prod_{i=1}^{k}\left[\frac{Q(t_2)/X_i(t_2)}{Q(t_1)/X_i(t_1)}\right]^{\bar{s}_i}. \tag{8}$$

$$\frac{f[X(t_2)]}{f[X(t_1)]} = \prod_{i=1}^{k}\left[\frac{X_i(t_2)}{X_i(t_1)}\right]^{1/2[s_i(t_2)\,+\,s_i(t_1)]},$$

which is a Törnqvist index of $X(t)$. Here, $s_i(t)$ is the share in costs of $X_i(t)$.

Rearranging equation (8) indicates that the change in unit costs can be expressed as

$$\frac{c^*[t_2,W(t_2)]}{c^*[t_1,W(t_1)]} = \tilde{c}[W(t_2),W(t_1)] \Big/ \prod_{i=1}^{k} \left[\frac{Q(t_2)/X_i(t_2)}{Q(t_1)/X_i(t_1)}\right]^{\bar{s}_i}. \tag{9}$$

Substituting equations (7) and (9) into equation (6) provides one scheme for decomposing product price changes into input, efficiency, and markup components:

$$\frac{P(t_2)}{P(t_1)} = \frac{M(t_2)}{M(t_1)} \cdot \frac{A(t_1)}{A(t_2)} \cdot \tilde{c}[W(t_2),W(t_1)]. \tag{10}$$

If one uses equation (4) to measure productivity growth, a slightly different accounting scheme emerges. In this case, one presumes c is translog, so $c[W(t_2)]/c[W(t_1)]$ equals a Törnqvist index of input prices:

$$\frac{c[W(t_2)]}{c[W(t_1)]} = \prod_{i=1}^{k} \left[\frac{W_i(t_2)}{W_i(t_1)}\right]^{\bar{s}_i}. \tag{11}$$

Equation (4) can then be used to implicitly define an efficiency index:

$$\tilde{A}(t_2,t_1) = \prod_{i=1}^{k} \left[\frac{W_i(t_2)}{W_i(t_1)}\right]^{\bar{s}_i} \Big/ \frac{c^*[t_2,W(t_2)]}{c^*[t_1,W(t_1)]}. \tag{12}$$

Analogously to equation (9), one can decompose unit cost changes as:

$$\frac{c^*[t_2,W(t_2)]}{c^*[t_1,W(t_1)]} = \prod_{i=1}^{k} \left[\frac{W_i(t_2)}{W_i(t_1)}\right]^{\bar{s}_i} \Big/ \tilde{A}(t_2,t_1). \tag{13}$$

Substituting equation (13) into equation (6) provides the second price accounting scheme:

$$\frac{P(t_2)}{P(t_1)} = \frac{M(t_2)}{M(t_1)} \prod_{i=1}^{k} \left[\frac{W_i(t_2)}{W_i(t_1)}\right]^{\bar{s}_i} \Big/ \tilde{A}(t_2,t_1). \tag{14}$$

Equations (10) and (14) are each internally consistent but differ in the presumptions they make about the way the data were generated. In equation (10) the presumption is that the data were generated by minimizing costs subject to a translog production function. In equation (14) the presumption is that costs were minimized subject to the production function which generates a translog cost function. Since a translog production function does not yield a translog cost function (and conversely), different accounting schemes follow from the different assumptions. However, since translog production and cost functions both provide second-order approximations to the technology, one would expect the two accounting schemes to give similar results. That, indeed, is true in the case of steel rails.

III

Table 1 presents details of the costs of producing steel rails in the United States in 1879, 1889, 1902, and 1910. The figures for the latter two years are derived from the U.S. Commissioner of Corporations (1913) investigation into costs, prices, and profits in the steel industry. The commissioner obtained and published details of the cost of producing most steel products made in the United States between 1902 and 1910. The figures for 1902 are averages for firms producing virtually all of the heavy steel rails made in America in that year. The 1910 figures are averages for all U.S. Steel's Bessemer rail mills. These mills produced 60.2 percent of American production in that year. The cost figures for 1901 and 1903–9 that the bureau reports are quite similar to the 1902 and 1910 figures, so conclusions founded on those years are indicative of conditions throughout the first decade following the U.S. Steel merger.[6]

The cost figures for 1879 and 1889 are less certain than those for 1902 and 1910. The 1889 figures are based on the U.S. Commissioner of Labor's (1891) survey of steelmaking costs in the United States, Great Britain, and the European continent. Although the cost estimates were based on the cost accounts of steel firms, the coverage was not as broad as the U.S. Bureau of Corporations survey. Moreover, the U.S. Commissioner of Labor (1891, p. 163) *Report* expressed concern that some of the leading American rail mills had refused to supply cost information. The 1879 figures are also from an international comparison of production costs. This survey was conducted by Alexander L. Holley, who was the leading American steelworks engineer. The survey was published privately by the Bessemer Association, which owned the American rights to the Bessemer patents, for the use of its licensees. Holley visited several European plants in 1879 and reported their costs. For comparison, he provided "representative costs" in the United States. It is the latter costs which are used in table 1. Given Holley's knowledge, the audience he was writing for, and the confidential nature of the communication, one presumes the American cost figures are indeed representative.[7]

[6] See U.S. Commissioner of Corporations (1913, pp. 209–10, 461) for 1910 rail production figures and pp. 143–49, 209–16, 346–68 for information on costs in the other years.

[7] It should be noted that none of these sources reports the quantity and price of all inputs consumed. That practice was followed in the case of metallic inputs, but in the cases of fuel, labor, miscellaneous materials, and capital, only the expense per ton was recorded. It was necessary to estimate input prices and implicitly input quantities from other sources. Those sources are discussed in the notes to table 1. It must also be stressed that all calculations reported in this paper assume that the costs in table 1 are minimum long-run average total costs.

TABLE 1

RAIL COSTS, PRICES, AND MARKUPS

Input	1879 W_i	1879 X_i/Q	1879 W_iX_i/Q	1889 W_i	1889 X_i/Q	1889 W_iX_i/Q	1902 W_i	1902 X_i/Q	1902 W_iX_i/Q	1910 W_i	1910 X_i/Q	1910 W_iX_i/Q
Metal	20.60	1.29	26.47	15.21	1.24	18.86	14.78	1.09	16.18	15.36	1.09	16.67
Fuel	1.06	2.08	2.20	1.04	2.04	2.13	1.68	.54	.91	1.55	.50	.77
Labor	.432	8.472	3.66	.577	5.806	3.35	.609	3.284	2.00	.697	2.396	1.67
Materials	.90	1.90	1.77	.81	1.85	1.50	.859	1.69	1.45	1.027	1.09	1.12
Capital	7.1848	.1851	1.33	5.8252	.2263	1.32	5.7942	.2675	1.55	6.2888	.2067	1.30
Average total cost	...	...	35.37	...	...	27.16	...	...	22.09	...	...	21.53
Price of rails	...	...	40.18	...	...	29.25	...	...	28.00	...	...	28.00
M	...	...	1.14	...	...	1.08	...	...	1.27	...	...	1.30

SOURCES.—Except as noted subsequently, the price (dollars per long ton) and quantity (long tons) consumed of metal (which includes manganese as well as pig iron and any scrap) and the expenses of the various inputs per ton of rails were computed from Holley (1881), U.S. Commissioner of Labor (1891, p. 173), and U.S. Commissioner of Corporations (1913, pp. 143–49, 209–16, 426–29, 460–68). In all cases it was necessary to combine separately reported costs for the production of ingots, the rolling of billets, and the rolling of rails. It was also necessary to estimate the capital cost component of the pig iron price for 1889. This cost was estimated as 10% of the capital invested in American blast furnaces in 1889 divided by the production of pig iron. These figures were taken from U.S. Census Office (1902, p. 28). Expenses for other inputs were broken into prices and quantities as follows. *Fuel:* The price of fuel (dollars per long ton) was taken to be the price of bituminous coal purchased by steelworks and rolling mills. For 1879 the price of coal is the price paid by Pennsylvania steelworks as given in U.S. Census Office (1883, p. 757) deflated by the ratio of the 1879 to 1880 value of the fuel and lighting component of the Warren and Pearson price index given in U.S. Bureau of the Census (1975, 1:201). The low price of metallic inputs reported by Holley and the dates at which he visited European works suggest he was reporting early 1879 American values. In mid-1879 American steel prices and steel input prices rose, and this rise is reflected in the 1879 census figures which were collected for the year June 1, 1879–May 31, 1880. Hence, the fuel price in the census was deflated. The U.S. Commissioner of Labor (1891) reported the quantity of coal consumed per ton of rails. The price of coal was taken to equal coal expense per ton of rails divided by fuel consumption. For 1902 and 1910 the price of fuel was taken to equal the price of coal consumed by steelworks and rolling mills in 1899 (given in U.S. Census Office 1902, p. 57), inflated in proportion to the rise in the fuel and lighting component of the BLS wholesale price index given in U.S. Bureau of the Census (1975, 1:200). For 1902 and 1910 fuel expenses in table 1 include the cost of steam. The quantity of fuel (in long tons) consumed equals fuel expense divided by the price of fuel. *Labor:* The price of labor (thousands of dollars per man-year) is the average earnings of workers employed in steelworks and rolling mills. For 1879 and 1889 the average is computed from U.S. Bureau of the Census (1913, p. 228). For 1902 and 1910 the average is taken from Douglas (1930, p. 271). The quantity of labor equals labor expense divided by the price of labor. *Materials:* The price of materials is taken to be the Warren and Pearson price index extended past 1890 by the BLS wholesale price index (U.S. Bureau of the Census 1975, 1:200–201). These have been divided by 100. The quantity of materials equals material expense divided by the price of materials. *Capital:* The price of capital equals the deflator for capital in the iron and steel industry developed by Creamer, Dobrovolsky, and Borenstein (1960, p. 260) multiplied by an interest rate plus a depreciation rate. The deflators given for 1880, 1890, 1900, and 1909 were applied to the years 1879, 1889, 1902, and 1910, respectively. The depreciation rate was presumed to be .059 on the basis of Creamer et al. (1960, p. 223). The interest rate was taken to be the Cowles Commission yield on industrial common stock as given by U.S. Bureau of the Census (1975, 2:1003). The quantity of capital equals capital expense divided by the price of capital. The U.S. Commissioner of Labor (1891) did not report any capital cost information, so the quantity of capital in 1889 was estimated as the average of the quantities of capital in 1879 and 1902. Capital expense in 1889 then equals that quantity multiplied by the 1889 price of capital. *Price of rails:* The price of rails was taken from U.S. Commissioner of Labor (1891, p. 179) and Temin (1964, pp. 284–85). For reasons given in the discussion of fuel prices, an early 1879 price of pig iron ($41 per ton) was chosen. The $40.18 shown in the table equals $41 less the royalty of $.82 per ton of rails ($.75 per ton of ingots). The royalty lapsed by 1889, and subtracting it from the price in 1879 seemed the easiest way to incorporate it.

NOTE.—W_i = the price of input i, X_i/Q = the consumption of input i per ton of rails, and W_iX_i/Q = the expense per ton of rails of input i. Average total cost equals the sum of W_iX_i/Q for each year. M = the price of rails divided by the average cost.

Between the 1880s and early 1900s the American steel industry became internationally competitive because the price of steel in the United States fell sharply. Table 1 and tables 2 and 3, which report the index numbers discussed in the previous section, make clear the decline in price was associated with a sharp decline in steelmaking costs. According to table 2, steelmaking costs in 1889 equaled 77 percent of costs in 1879. In 1902 costs were 81 percent of their 1889 level. The decline in costs subsequently slowed, for costs in 1910 equalled 97 percent of their 1902 level.

Equations (9) and (13) provide schemes for decomposing the cost changes into efficiency and input price components. The results from using both equations are exhibited in table 2, and, as the reader can verify, the results are nearly identical. Between 1879 and 1889 falling input prices were the main cause of falling costs. Thereafter input prices rose slowly. Between 1889 and 1902 costs continued to fall because productivity growth was rapid. After 1902, the rate of productivity growth slackened and input prices rose, so the decrease in costs between 1902 and 1910 was meager.

Equation (7) indicates that productivity increased at the rate of 0.6 percent per year between 1879 and 1889. Between 1889 and 1902 the rate increased to 1.6 percent and then declined to 0.8 percent in the period 1902–10. (Eq. [12] gives essentially the same results.) Throughout the period 1879–1910, the average product of labor continuously rose, and the average product of capital generally fell. Since the ratio of the wage rate to the rental price of capital consistently increased, the change in the average products of labor and capital might have been due to factor substitution. Moreover, recent econometric work suggests that technical change was labor augmenting in this period (Cain and Paterson 1979). Technical change and factor substitution reinforced each other in determining the labor-capital mix.

Productivity growth was highest between 1889 and 1902. What distinguished those years from the others was the marked rise in the average products of fuel and metallic inputs. Fuel consumption declined as the mixer (invented in 1889) was adopted by American steelworks. In earlier years, it had been impossible to use the molten pig iron from blast furnaces directly in Bessemer converters. The pig iron was cast and then remelted. By 1902 most Bessemer steel was made from molten iron conveyed directly from the blast furnace. The saving in fuel was impressive. The productivity of metallic inputs also increased due to developments in other areas of the steel mill, in this case due to the growth of open hearth steelmaking capacity. Bessemer shops and rolling mills generated a large volume of scrap in the form of poor rolls and sheared ends. These could not be recycled in

TABLE 2

DECOMPOSING UNIT COST CHANGES

PERIOD	By Eq. (9)					By Eq. (13)				
	$\frac{c^*[t_2,W(t_2)]}{c^*[t_1,W(t_1)]}$	=	$\frac{A(t_1)}{A(t_2)}$	·	$\tilde{c}[W(t_2),W(t_1)]$	$\frac{c^*[t_2,W(t_2)]}{c^*[t_1,W(t_1)]}$	=	$\frac{1}{\tilde{A}(t_2,t_1)}$	·	$\frac{c[W(t_2)]}{c[W(t_1)]}$
1879–89	.7679	=	(.9367)	·	(.8198)	.7679	=	(.9397)	·	(.8172)
1889–1902	.8133	=	(.7962)	·	(1.0215)	.8133	=	(.7994)	·	(1.0174)
1902–10	.9746	=	(.9305)	·	(1.0475)	.9746	=	(.9242)	·	(1.0545)

SOURCE.—Table 1: $c^*[t_2,W(t_2)]/c^*[t_1,W(t_1)]$ is observed unit cost (from table 1) in the terminal year of the period divided by observed unit cost in the initial year. $A(t_1)/A(t_2)$ was computed according to eq. (7); $\tilde{c}[W(t_2),W(t_1)]$ according to eq. (8); $c[W(t_2)]/c[W(t_1)]$ according to eq. (11); and $\tilde{A}(t_2,t_1)$ according to eq. (12).

TABLE 3

DECOMPOSING PRICE CHANGES

PERIOD	By Eq. (10)							By Eq. (4)						
	$\frac{P(t_2)}{P(t_1)}$	=	$\frac{M(t_2)}{M(t_1)}$	·	$\frac{A(t_1)}{A(t_2)}$	·	$\tilde{c}[W(t_2),W(t_1)]$	$\frac{P(t_2)}{P(t_1)}$	=	$\frac{M(t_2)}{M(t_1)}$	·	$\frac{1}{\tilde{A}(t_2,t_1)}$	·	$\frac{c[W(t_2)]}{c[W(t_1)]}$
1879–89	.7280	=	(.9480)	·	(.9367)	·	(.8198)	.7280	=	(.9480)	·	(.9397)	·	(.8172)
1889–1902	.9573	=	(1.1771)	·	(.7962)	·	(1.0215)	.9573	=	(1.1771)	·	(.7994)	·	(1.0174)
1902–10	1.000	=	(1.0261)	·	(.9305)	·	(1.0475)	1.000	=	(1.0261)	·	(.9242)	·	(1.0545)

SOURCE.—Table 1: $P(t_2)/P(t_1)$ is the ratio of the price in the terminal year of the period to the price in the initial year. See table 2 for the definitions of other variables.

Bessemer converters but could be remelted in open hearth furnaces. In 1889 only 11 percent of American steel was open hearth, but by 1902 38 percent was open hearth (Temin 1964, pp. 272–73). The development of this process allowed the recycling of scrap. Tables 1 and 2 follow the nineteenth-century accounting tradition of charging Bessemer rail mills for their net consumption of metallic inputs. The recycling of Bessemer scrap in open hearth plants thus shows up as a rise in the productivity of metallic inputs in Bessemer rail production.

Table 1 indicates that the fall in production costs was not entirely passed on to consumers in lower prices. An apparent consequence of the U.S. Steel merger was to raise markups to the 30 percent range for at least the first decade after the merger. In 1879 and 1889 markups had been much lower.

Table 3 uses equations (10) and (14) to decompose price changes into markup, efficiency, and input price components. The price of rails in 1889 equalled 72.8 percent of the price in 1879. Some of the decline in price was due to a decline in the markup; however, most of the decline can be ascribed to falling costs. As indicated previously, declining input prices were the predominant cause of declining costs. The fall in input prices was also decisive for the decline in price.

Between 1889 and 1902, costs declined 19 percent but rail prices declined only 4 percent. The discrepancy arose because markups rose 18 percent. This result indicates a major consequence of the U.S. Steel merger. Steel prices in the early twentieth century were set at levels prevailing about 1890, and the steel producers absorbed all of the decline in costs (due to technical progress in the case of rails) as excess profits.

Table 3 indicates that there was little change in price, costs, or the determinants of costs between 1902 and 1910. The steel industry was sufficiently collusive so that a markup of 30 percent over unit costs could be maintained for a decade.

IV

These conclusions regarding productivity growth and the market power exercised after the U.S. Steel merger are at variance with recent conclusions advanced by McCloskey (1973, p. 99) and Temin (1964, pp. 186–93, 218–20). Both noted that the ratio of the price of steel rails to the price of pig iron was trendless from the 1880s to the First World War. (This constancy is illustrated by the data in table 1.) Both inferred from these price data that productivity growth ceased in the 1880s, and Temin also inferred that market control after the U.S. Steel merger was not effective in raising the price of rails.

To choose between McCloskey's and Temin's conclusions and those

advanced here, it is necessary to scrutinize the theoretical underpinnings of the calculations employed. McCloskey's procedure consists of the (approximate) application of an identity deduced by Jorgenson (1966, pp. 2–4). Although the assumption is not obviously apparent in the derivation, it is the case that this procedure only correctly measures productivity growth (and, hence, accounts for price changes) for industries in competitive long-run equilibrium. This will be shown subsequently. It is worth considering the matter thoroughly, both to resolve the issue at hand and because Jorgenson's identity has been widely used by economic historians to measure productivity growth and account for price changes.[8]

Jorgenson begins with the accounting identity that a firm's revenues necessarily equal the total income earned by the firm's inputs. In the case of a single product firm,

$$PQ = \sum_{i=1}^{k} W_i^* X_i, \tag{15}$$

where W_i^* are the accounting valuations of the inputs X_i. Totally differentiating equation (15) with respect to time, dividing the result by equation (15), and regrouping terms gives:

$$\frac{\dot{Q}}{Q} - \sum_{i=1}^{k} s_i^* \frac{\dot{X}_i}{X_i} = \sum_{i=1}^{k} s_i^* \frac{\dot{W}_i^*}{W_i} - \frac{\dot{P}}{P}. \tag{16}$$

Here the shares are $s_i^* = W_i^* X_i / \Sigma W_i^* X_i$.

Jorgenson defined productivity to equal the ratio of output to total input: $E \equiv Q/X$. Totally differentiating this expression with respect to time implies that the rate of change of efficiency equals the rate of change of output less the rate of change of total input: $\dot{E}/E = \dot{Q}/Q - \dot{X}/X$. Defining $\dot{X}/X$ to equal $\Sigma s_i^*(\dot{X}_i/X_i)$ implies

$$\frac{\dot{E}}{E} = \frac{\dot{Q}}{Q} - \sum_{i=1}^{k} s_i^* \frac{\dot{X}_i}{X_i} = \sum_{i=1}^{k} s_i^* \frac{\dot{W}_i}{W_i} - \frac{\dot{P}_i}{P_i}. \tag{17}$$

Jorgenson contended, moreover, that Solow (1957) deduced the left-hand equality from a production model.

If correct, equation (17) has an important implication for produc-

[8] North (1968, p. 954), Shepherd and Walton (1972, pp. 60–63), Harley (1973, p. 376), and McCloskey (1973, p. 86) cite Jorgenson (1966). Sandberg (1974, p. 100) cites the thesis version of McCloskey (1973) for a proof of the equivalence of the G. T. Jones index of real cost and the right-hand side of eq. (16) of this paper and goes on to assert the equivalence of that index and Solow's (1957) index. Most of these economic historians recognize that the identity is not a sound basis for measuring productivity change when long-run competitive equilibrium does not prevail, and a number of ad hoc adjustments are developed to deal with the problem. Lindert and Trace (1971) do develop productivity indices that exclude excess profits.

tivity measurement, namely, that productivity growth can be inferred from the growth of input prices less the growth in output price *without any explicit account being taken of markups.* This result justified McCloskey's procedure and appears inconsistent with equations (10) and (14) in this paper.

Equation (17) is an identity and hence always true. The difficulty with using it as the basis of productivity measurement is that $\dot{E}/E$ equals the rate of productivity growth in the Solow (1957) sense only if the firm (or industry) whose productivity growth is being measured is in long-run, competitive equilibrium. This condition was not made explicit in Jorgenson's derivation. To see the importance of the competitive equilibrium assumption, recall Solow's derivation. Solow (1957) assumes a production function of the form of equation (1) and identifies the rate of productivity growth with $\dot{A}/A$. Totally differentiating equation (1) with respect to time, dividing the result by equation (1), and rearranging allows $\dot{A}/A$ to be expressed as

$$\frac{\dot{A}}{A} = \frac{\dot{Q}}{Q} - \sum_{i=1}^{k} \frac{(\partial f/\partial X_i)X_i}{f(X)} \cdot \frac{\dot{X}_i}{X_i}, \tag{18}$$

where the time symbol t is suppressed for simplicity. If the firm faces exogenous input prices and minimizes costs, then the necessary conditions for cost minimization can be rewritten as $W_i/\Sigma W_i X_i = (\partial f/\partial X_i)/f(X)$. In that case equation (18) can be rewritten as

$$\frac{\dot{A}}{A} = \frac{\dot{Q}}{Q} - \sum_{i=1}^{k} s_i \frac{\dot{X}_i}{X_i}, \tag{19}$$

where $s_i = W_i X_i/\Sigma W_i X_i$.

Equation (19) and the left-hand equality of equation (17) are almost identical. They would be identical if $s_i^* = s_i$. Since equation (17) is an accounting identity, one must make some assumptions about the institutional and accounting framework in order to determine if $s_i^* = s_i$. Reasonable, simple assumptions are that inputs are available to the firm in infinitely elastic supply at prices W_i, any difference between revenues and the opportunity cost of inputs accrues as excess profit (or loss) to capital, and all inputs besides capital are valued in the firm's accounts at their supply prices W_i. Given these presumptions $W_i^* = W_i$ for all inputs besides capital. For capital (which will be indexed as the kth input), $W_k^* \gtreqqless W_k$ if revenues exceed, equal, or fall short of the opportunity cost of the inputs.

If the firm is in competitive, long-run equilibrium, revenues equal the opportunity cost of all the inputs and $W_k^* = W_k$. In this case $s_i^* = s_i$, and equation (17) correctly measures the rate of productivity growth $\dot{A}/A$. Moreover, in this case, productivity growth can be inferred from

input and product price movements without taking markups into account. This result is consistent with the framework of Section II of this paper, for, if revenues always equal opportunity costs, $M(t_1) = M(t_2) = 1$ and the markup terms disappear from equation (14). Efficiency change can then be inferred from changes in product and input prices.

Suppose, on the other hand, that the markup exceeds one. In that case revenues exceed opportunity costs, $W_k^* > W_k$, and $\Sigma W_i^* X_i > \Sigma W_i X_i$. Hence, $s_k^* > s_k$ and $s_i^* < s_i$ for all inputs besides capital. Unless all inputs (including capital) grow at the same rate $\Sigma s_i^*(\dot{X}_i/X_i) \neq \Sigma s_i(\dot{X}_i/X_i)$ and the rate of productivity growth cannot be inferred from equation (17). The same conclusion follows if markups are less than one. Therefore, if markups do not equal one, equation (17) cannot be used to measure the rate of shift of the production function. In particular, one cannot infer the rate of productivity growth from differences in the rates of growth of input and product prices.

Substantial errors result if equation (17) is used when markups do not equal one. The errors are apparent if equation (17) is applied to the data in table 1. That application requires that some discrete approximation be adopted. The most useful approximation is to approximate the time-varying shares s_i^* by the average of the shares prevailing at times t_1 and t_2, $\bar{s}_i^* = 1/2\,[s_i^*(t_1) + s_i^*(t_2)]$, and integrate the left-hand equality in equation (17) between t_1 and t_2. Exponentiating the integral yields:

$$\frac{E(t_2)}{E(t_1)} = \prod_{i=1}^{k} \left[\frac{Q(t_2)/X_i(t_2)}{Q(t_1)/X_i(t_1)} \right]^{\bar{s}_i^*}. \tag{20}$$

If one treats equation (19) analogously, the result is equation (7). Again the only difference between the putative productivity measures is in the shares. Numerically, however, the differences are not inconsequential. For the periods 1879–89, 1889–1902, and 1902–10 $A(t_2)/A(t_1)$ equals 1.0676, 1.2560, and 1.0747, respectively. For the same periods $E(t_2)/E(t_1)$ equals 1.0408, 1.1904, and 1.1199. The term $E(t_2)/E(t_1)$ understates the average annual rate of productivity growth by 40 percent in 1879–89 and 24 percent in 1889–1902 and overstates the rate by 59 percent in 1902–10. Assuming long-run, competitive equilibrium when it does not obtain leads to serious errors in productivity measurement.

V

In this paper it has been shown that American rail prices fell in the late nineteenth century because production costs declined sharply. Two unexpected findings include the discovery that the rate of pro-

ductivity growth remained high into the twentieth century and the discovery that rail producers managed to maintain steel prices at a level 30 percent above average total cost for a decade after the U.S. Steel merger. Temin's (1964, pp. 189–93) recent work on the merger concluded that the markets for rolled products remained competitive after the merger. That conclusion is certainly not true for rails. Similar evidence for other products suggests that markups were generally high.[9] A reassessment of the significance of the U.S. Steel merger in this regard seems in order.

In order to establish these conclusions, it was necessary to develop a more elaborate price accounting scheme than those previously available. As part of that effort, considerable attention was devoted to the circumstances under which one could infer the rate of productivity change from the difference in the rates of input price and product price changes. A correct inference can be made only if the industry is in competitive long-run equilibrium. If that condition does not obtain, serious errors are likely to result. These difficulties can be avoided if productivity growth is measured with equation (7) or (12).

References

Allen, Robert C. "Accounting for Price Changes." Discussion Paper no. 78-14, Univ. British Columbia, Dept. Econ., 1978.

———. "International Competition in Iron and Steel, 1850–1913." *J. Econ. Hist.* 39 (December 1979): 911–37.

Cain, L. P., and Paterson, D. G. "Factor Biases and Technical Change in Manufacturing: The American System, 1850." Mimeographed, Univ. British Columbia, Dept. Econ., 1979.

Creamer, Daniel; Dobrovolsky, Sergei P.; and Borenstein, Israel. *Capital in Manufacturing and Mining: Its Formation and Financing.* Princeton, N.J.: Princeton Univ. Press, 1960.

Diewert, W. E. "Exact and Superlative Index Numbers." *J. Econometrics* 4 (May 1976): 115–45.

———. "Aggregation Problems in the Measurement of Capital." In *The Measurement of Capital,* edited by Dan Usher. Chicago: Univ. Chicago Press (for Nat. Bur. Econ. Res.), 1980.

Douglas, Paul H. *Real Wages in the United States, 1890–1926.* Boston: Houghton Mifflin, 1930.

Harley, C. K. "On the Persistence of Old Techniques: The Case of North American Wooden Shipbuilding." *J. Econ. Hist.* 33 (June 1973): 372–98.

Holley, Alexander L. *Holley's Reports to the Bessemer Steel Co. Limited: Bessemer and Rolling Mill Practice and Costs at the West Cumberland Iron & Steel Works and at Brown, Bailey & Dixon's Works.* New York: Russell Bros., 1881.

[9] U.S. Commissioner of Corporations (1913, pp. 520–31) shows substantial profits, but the figures as presented are not directly comparable with those in this paper and do not throw light on excess profits in the rolled product stage of the industry. Preliminary reworking of these figures using the methods developed here shows high markups for most rolled products.

Jorgenson, Dale W. "The Embodiment Hypothesis." *J.P.E.* 74 (February 1966): 1–17.

Lindert, Peter H., and Trace, Keith. "Yardsticks for Victorian Entrepreneurs." In *Essays on a Mature Economy: Britain after 1840*, edited by Donald N. McCloskey. Princeton, N.J.: Princeton Univ. Press, 1971.

McCloskey, Donald N. *Economic Maturity and Entrepreneurial Decline: British Iron and Steel, 1870–1913*. Cambridge, Mass.: Harvard Univ. Press, 1973.

North, Douglass C. "Sources of Productivity Change in Ocean Shipping, 1600–1850." *J.P.E.* 76 (September/October 1968): 953–70.

Sandberg, Lars G. *Lancashire in Decline: A Study in Entrepreneurship, Technology, and International Trade*. Columbus: Ohio State Univ. Press, 1974.

Shepherd, James F., and Walton, Gary M. *Shipping, Maritime Trade, and the Economic Development of Colonial North America*. Cambridge: Cambridge Univ. Press, 1972.

Solow, Robert M. "Technical Change and the Aggregate Production Function." *Rev. Econ. and Statis.* 39 (August 1957): 312–20.

Temin, Peter. *Iron and Steel in Nineteenth-Century America: An Economic Inquiry*. Cambridge, Mass.: MIT Press, 1964.

U.S. Bureau of the Census. *Thirteenth Census of the United States*. Vol. 10. *Manufactures: Reports for Principal Industries*. Washington: Government Printing Office, 1913.

———. *Historical Statistics of the United States*. Washington: Government Printing Office, 1975.

U.S. Census Office. *Report on the Manufactures of the United States at the Tenth Census*. Washington: Government Printing Office, 1883.

——. *Twelfth Census of the United States*. Vol. 10. *Manufactures: Part IV, Special Reports on Selected Industries*. Washington: Government Printing Office, 1902.

U.S. Commissioner of Corporations. *Report on the Steel Industry*. Pt. 3, *Cost of Production*. Washington: Government Printing Office, 1913.

U.S. Commissioner of Labor. *Sixth Annual Report: Cost of Production: Iron, Steel, Coal, etc.* Rev. ed. Washington: Government Printing Office, 1891.

Some Evidence on Cross-Sector Effects of the Minimum Wage

George E. Tauchen

Duke University

This paper tests Mincer's minimum-wage model by estimating reduced-form wage and employment equations for both the covered and uncovered sectors in nine regions of the United States. As theory predicts, in regions with comparatively small covered-sector demand elasticities, the northern and midwestern regions, the uncovered-sector wage increases after a minimum-wage hike; and in regions with comparatively large demand elasticities, the southern and western regions, the uncovered-sector wage decreases. Because of data limitations the uncovered-sector employment effect could not be estimated sharply, and so its relationship to the covered-sector demand elasticity is weak.

I. Introduction

Economic theory predicts that a minimum-wage hike decreases employment of low-skill workers in the covered sector. Theory also suggests that a minimum-wage hike affects the equilibrium wage and the level of employment in the uncovered sector though, as pointed out by Mincer (1976), the directions of the uncovered-sector effects are ambiguous. On the one hand, if the absolute labor demand elasticity is relatively small so that firms eliminate only a few covered jobs after the wage hike, then workers find the minimum wage attractive enough to justify the risk of unemployment while searching for jobs in the covered sector. In this case labor leaves the uncovered sector, and the wage thereby increases. On the other hand, if the

I am grateful to Christopher A. Sims, Helen Tauchen, and an anonymous referee for many helpful comments and suggestions.

[*Journal of Political Economy*, 1981, vol. 89, no. 3]

demand elasticity is relatively large so that the return to search is low, then displaced workers enter the uncovered sector, and the wage decreases.

The objective of this paper is to test Mincer's characterization of the effects of a minimum-wage hike on the uncovered sector. In the empirical work the covered sector is an aggregation of four low-wage nondurable manufacturing industries; the uncovered sector is agriculture prior to coverage under the Fair Labor Standards Act. Time-series data are used to estimate the covered-sector labor demand elasticity and the uncovered-sector wage and employment effects of a minimum-wage hike for nine regions of the United States. As theory predicts, in regions with comparatively small demand elasticities, the northern and midwestern regions, the uncovered-sector wage increases after a minimum-wage hike; and in regions with comparatively large demand elasticities, the southern and western regions, the uncovered-sector wage correspondingly decreases. The uncovered-sector employment effect is not as sharply estimated as the wage effect, and so its relationship to the demand elasticity is weak. This result is explained by limitations of the data.

There is a potentially serious problem with using regional time-series data to test the minimum-wage model. The problem arises because it is impossible to get long time series of data on the level of employment of persons earning exactly the minimum wage. Although average hourly earnings in the four low-wage industries are among the lowest in the economy, they are still well above the minimum wage, indicating that the data series contain contributions from high-wage labor. Thus the dependent variables in the regressions are total employment and the average wage, both aggregated over the wage distribution. It is conceivable, then, that the observed regional variation in the covered-sector demand elasticity is a statistical artifact reflecting regional differences in the mix of low- and high-wage labor. In other words, the true demand elasticity for minimum-wage grade labor could be the same in all regions, while the estimated demand elasticity is larger in the South simply because there are more low-wage workers there.

The evidence presented in subsequent sections suggests that the spurious aggregation effect is small. Briefly, the argument is as follows. Think of low- and high-wage labor as imperfect substitutes in a production function with constant returns to scale. Now let the minimum wage increase by 1 percent. There are both scale (output) and substitution effects. The scale effect causes employment of both factors, and hence total employment, to decline equiproportionately. The substitution effect, on the other hand, causes low-wage employment to decline and high-wage employment to increase. Intuitively, and as is shown in Appendix A, the observed proportionate increase

in the average wage ($\Delta\overline{W}/\overline{W}$) is no less than the share (s_1) of low-wage labor in total labor costs; if the substitution effect is zero, then $\Delta\overline{W}/\overline{W}$ equals s_1. The regional estimates of $\Delta\overline{W}/\overline{W}$ from the regressions turn out to be very close to regional estimates of s_1 derived from independent wage distribution data. This can happen only if the major impact of a minimum-wage hike is the scale effect and not the substitution effect. Since the scale effect dominates, the observed proportionate decline in total employment must be close to the proportionate decline in low-wage employment; that is, the aggregation bias is small in the covered sector.

The importance of the scale effect together with the cost-share data can explain why there are regional differences in the demand elasticity. The size of the scale effect depends upon the increase in the marginal cost of production caused by a hike in the minimum wage. Since s_1 is 0.14 in the North and 0.32 in the South, the proportionate increase in marginal cost is larger in the South, and so output and total employment contract proportionately more there. This is consistent with the idea that one purpose of the national minimum wage is to reduce the cost advantage of firms located in low-wage regions of the country. It is also consistent with Silberman and Durden's (1976, p. 325) finding that Southern congressmen are less likely than other congressmen to vote in favor of a minimum-wage hike.

The remainder of this paper is organized as follows. Section II describes a simple two-sector labor market model which is used to guide the empirical work. Sections III and IV report the empirical findings for the covered and uncovered sectors. Section V uses the theoretical model to interpret the results. Section VI contains the concluding remarks.

II. Two-Sector Model

The following two-sector model of the low-wage labor market is used to illustrate the restrictions being tested and to guide the empirical work. The model consists of two labor demand schedules (one for each sector), an equilibrium condition connecting the minimum wage and the competitively determined uncovered-sector wage, and an aggregate labor supply relationship. Formally, the model is:

$$\text{labor demand (covered):}\quad N^c = D^c(W/W_2, X^c),\quad D_1^c < 0, \tag{1a}$$

$$\text{labor demand (uncovered):}\quad N^u = D^u(W^u/W_2, X^u),\quad D_1^u < 0, \tag{1b}$$

$$\text{equilibrium condition:}\quad W^u = \left(\frac{\delta N^c}{U + \delta N^c}\right) W,\quad \delta > 0, \tag{1c}$$

$$\text{labor supply:}\quad N^c + N^u + U = S(W^u/P),\quad S' > 0, \tag{1d}$$

where N^c, N^u = employment in the covered (c) and uncovered (u) sector; W = minimum wage; W^u = wage in the uncovered sector; W_2 = control wage, $W_2 > W$; X^c, X^u = vectors of exogenous shift variables; U = unemployment; δ = turnover parameter; and P = general price level. The labor demand equation (1a) expresses employment in the covered sector, N^c, as a function of the relative minimum wage, W/W_2, where W_2 is an exogenously determined measure of high-skill wages, and as a function of exogenous shift variables, X^c. Similarly, equation (1b) expresses employment in the uncovered sector as a function of the uncovered-sector wage relative to the high-skill wage, W^u/W_2, and as a function of exogenous variables, X^u. Equation (1c) is from Mincer (1976). In equilibrium the uncovered-sector wage must equal the expected return from search in the covered sector (equal to the probability of getting a job times minimum wage). The parameter δ in (1c) is the proportion of covered-sector employees who each period must enter an employment lottery with the unemployed. Finally, equation (1d) relates the total labor force to the real wage in the uncovered sector.

The four endogenous variables in the system (1) are N^c, N^u, W^u, and U; the exogenous variables are W, W_2, P, X^c, and X^u. Note first that (1a) is the reduced-form expression for N^c. Now write the reduced form for the other three endogenous variables as

$$N^u = N^u(W, W_2, P, X^c, X^u), \tag{2a}$$

$$W^u = W^u(W, W_2, P, X^c, X^u), \tag{2b}$$

$$U = U(W, W_2, P, X^c, X^u). \tag{2c}$$

Theory places restrictions on the reduced-form equations. Specifically, Mincer shows

$$\frac{\partial W^u}{\partial W} \gtreqqless 0 \quad \text{as} \quad e^c \lesseqqgtr \delta, \tag{3a}$$

$$\frac{\partial N^u}{\partial W} \lesseqqgtr 0 \quad \text{as} \quad e^c \lesseqqgtr \delta, \tag{3b}$$

where e^c is the absolute elasticity of labor demand in the covered sector (the absolute elasticity of [1a] with respect to the minimum wage). The unemployment effect is positive,

$$\frac{\partial U}{\partial W} > 0, \tag{4}$$

and the sign of the labor force effect agrees with the sign of the uncovered-sector wage effect,

$$\frac{\partial N^c}{\partial W} + \frac{\partial N^u}{\partial W} + \frac{\partial U}{\partial W} \gtreqqless 0 \quad \text{as} \quad e^c \lesseqqgtr \delta. \tag{5}$$

Some of these results have been previously investigated empirically. There are, of course, many studies of the disemployment effect (1a) and of the unemployment effect (4). The Mincer study measures the labor force effect (5) for teenagers and young adults. After finding a negative labor force effect, Mincer infers $e^c > \delta$ and uses (3) to conclude that $\partial W^u/\partial W < 0$, that is, that, as a group, low-wage workers' prospects decline after a minimum-wage hike. Gardner's (1972) earlier work with aggregate U.S. agricultural data is consistent with Mincer's findings, though Gardner's estimates lack statistical significance at conventional levels. The purpose of this study is to use regional variation in the demand elasticity e^c to test the characterization (3a) and (3b). If there is small regional variation in the parameter δ, then regions with relatively small values of e^c should show $\partial W^u/\partial W > 0, \partial N^u/\partial W < 0$, and vice versa for regions with relatively large values of e^c.

III. The Covered Sector

The first step in the empirical work is to estimate the demand equation (1a) to get regional estimates of the covered-sector demand elasticity, e^c.[1] The data are annual observations on employment (1949–74) and average hourly earnings (1958–74) in an aggregation of four low-wage nondurable manufacturing industries: tobacco (SIC 21), textiles (SIC 22), apparel (SIC 23), and leather (SIC 31).[2] As mentioned in the Introduction, there is a problem because the employment variable includes workers for whom the legal minimum wage is economically ineffective. The effects of this skill-group aggregation can be characterized. Appendix A shows that regardless of the substitution pattern within aggregate labor, the observed (absolute) disemployment elasticity underestimates the sought-after demand

[1] The states within each region are: New England: Vt., Mass., N.H., Conn., R.I., Maine; Middle Atlantic: N.Y., Pa., N.J.; East North Central: Ohio, Wis., Ill., Ind., Mich.; West North Central: Mo., Iowa, N.Dak., S.Dak., Nebr., Kans., Minn.; South Atlantic: Va., Del., Ga., Md., W.Va., N.C., S.C., Fla.; East South Central: Ky., Tenn., Miss., Ala.; West South Central: Ark., La., Tex., Okla.; Mountain: Nev., Wyo., Mont., N.Mex., Utah, Idaho, Colo., Ariz.; Pacific: Calif., Oreg., Wash.

[2] Two-digit manufacturing industries were defined as low-wage industries if U.S. average hourly earnings in 1960 were no more than \$1.70; the legal minimum was upped from \$1.00 to \$1.15 in September 1961. Only the four indicated nondurable manufacturing industries met the criterion. No other two-digit nondurable manufacturing industry had an average wage below \$2.00 in 1960; the lowest average wage in 1960 of two-digit durable industries was \$1.88 in furniture (SIC 25). Annual observations on state employment and average hourly earnings were collected from U.S. Bureau of Labor Statistics (1977). In the case of missing values, the data from nearby larger states were spliced into the shorter series. This allowed development of usable state employment series from 1949 forward and wage series from 1958 forward. The series were aggregated over the four industries and over contiguous states to nine regions.

TABLE 1

COVERED-SECTOR EMPLOYMENT AND WAGE EQUATIONS: SIGNIFICANCE TESTS[a]

	Dependent Variable: n^c (1949–74) Significance Test $F(3,17)$		Dependent Variable: w^c (1958–74) Significance Test $F(3,8)$	
	w [−1 to +1]	w_2 [−1 to +1]	w [−1 to +1]	w_2 [−1 to +1]
United States	****	****	*	****
New England	...	****	...	****
Middle Atlantic	...	****	...	****
East North Central	****	****	...	****
West North Central	*	****	**	****
South Atlantic	****	****	**	****
East South Central	****	****	***	****
West South Central	****	****	****	****
Mountain	****	****	N.A.[b]	N.A.
Pacific	****	****	...	****

NOTE.—Dependent variables: n^c = log of total employment, four low-wage nondurable manufacturing industries, by region as indicated; w^c = log of average hourly earnings, four low-wage nondurable manufacturing industries, by region as indicated. Independent variables: w = log of minimum wage; w_2 = log of average hourly earnings, all manufacturing, excluding average hourly earnings in the four low-wage industries; z = log of the FRB index of industrial production, all manufacturing. Lags are in brackets; negative lags are leads (future terms).

[a] Test that indicated group of distributed lag coefficients is zero.

[b] No wage data are available for the Mountain region.

* 15%.

** 10%.

*** 5%.

**** 1%.

elasticity. In addition, the observed elasticity of the covered-sector wage with respect to the minimum wage overestimates the share of minimum-wage grade labor in total labor costs. Below, evidence will be presented that e^c is only slightly underestimated.

To focus attention on economics the less important technical details of the regressions are put in Appendix B. The results in tables 1 and 2 are from distributed lag regressions of the logs of covered-sector employment and average hourly earnings, n^c and w^c, on the log of the minimum wage, w, the control wage, w_2, and the cyclical index, z. The control wage is the log of average hourly earnings, all manufacturing, net of wages in the four low-wage industries; the cyclical variable is the log of the FRB index of industrial production.[3] Each regression

[3] w, w_2, and z are the logarithms of the annual average of monthly values. The level W_2 was computed as $W_2 = (EW - \Sigma_j E_j W_j)/(E - \Sigma_j E_j)$, where E is production worker employment and W is production worker average hourly earnings. Unsubscripted values refer to all manufacturing; the subscript j indexes the four low-wage industries. The basic data were taken from U.S. Bureau of Labor Statistics (1976) and from issues of *Employment and Earnings* (U.S. Bureau of Labor Statistics 1976–77). The industrial production index was taken from U.S. Department of Commerce (1976) and issues of the *Survey of Current Business* (U.S. Department of Commerce 1976–77).

TABLE 2

COVERED-SECTOR EMPLOYMENT AND WAGE EQUATIONS: SUMS OF DISTRIBUTED LAG COEFFICIENTS (Long-Run Effects)

	DEPENDENT VARIABLE: n^c			DEPENDENT VARIABLE: w^c		
	w [−1 to +1]	w_2 [−1 to +1]	z [0]	w [−1 to +1]	w_2 [−1 to +1]	z [0]
United States	−.2897	−.3124	.4370	.1865	.6936	−.0164
	(.0328)	(.0599)	(.0373)	(.0872)	(.1205)	(.0591)
New England	−.0394	−.9013	.5185	.1574	.5848	−.0836
	(.0847)	(.1468)	(.0794)	(.1148)	(.1590)	(.0777)
Middle Atlantic	−.1242	−.5710	.4159	.1406	.5827	−.1322
	(.0545)	(.0856)	(.0508)	(.1045)	(.1410)	(.0704)
East North Central	−.3812	−.3972	.5066	−.0099	.4943	−.2096
	(.0714)	(.1209)	(.0671)	(.1289)	(.1790)	(.0872)
West North Central	−.0987	.4048	.4261	.3500	.2371	−.2633
	(.0712)	(.0997)	(.0640)	(.1134)	(.1541)	(.0766)
South Atlantic	−.2237	−.2899	.4270	.1580	.8821	.1406
	(.0571)	(.0925)	(.0533)	(.0725)	(.0989)	(.0490)
East South Central	−.2042	−.2431	.5651	.3569	.6847	.0319
	(.0496)	(.0687)	(.0443)	(.0713)	(.0926)	(.0471)
West South Central	−.6360	−.0883	.6708	.4442	.6483	.0340
	(.0875)	(.1185)	(.0774)	(.1456)	(.1197)	(.0985)
Mountain	−.7589	.5059	1.4860	N.A.	N.A.	N.A.
	(.3909)	(.5911)	(.3606)			
Pacific	−.3984	1.0035	.5304	.2072	.4694	−.0685
	(.0746)	(.1178)	(.0694)	(.1366)	(.1899)	(.0924)

NOTE.—SE in parentheses. In the employment and wage equations df = 17, 8, respectively. See table 1 for the definitions of the variables. N.A. = data not available. Lags are in brackets.

included trend and a constant.[4] The regressions do not control for state minimum wages because in this sample state minimums are either at or below the federal minimum wage.[5]

The F-tests in table 1 indicate that the distributed lag coefficients on the minimum wage are significant as a group in nearly all of the equations. Now consider the long-run elasticities reported in table 2. As theory predicts, the sign of the minimum-wage effect is negative in all of the employment equations and positive in all but one wage equation. Moreover, the magnitude of the effect displays important regional variation. The disemployment effect appears to be negligible in the Northeast and upper Midwest, and it seems to be rather important through the South, the Southwest, and the Pacific coast. Also, the elasticities of employment and the average wage with respect to the minimum wage are consistent with one another, in the sense that, wherever the effect on the wage is small, so is the effect on employment.

To my knowledge, there are no other comparable published estimates of the disemployment effect by region or state. The estimated elasticity of -0.29 for the United States, however, is comparable with the findings of many other researchers. It is just outside Gramlich's (1976, p. 431) bounds of -0.05 to -0.25 for published estimates.[6]

Comparing these long-run effects to regional data on the wage distribution in the covered sector suggests that the major impact of the minimum wage is the scale or output effect. To see this, consider the 1964 wage distribution displayed in table 3. These data were

[4] Throughout this paper regional dependent variables are regressed on national variables. This is due to data limitations, but a strong case can be made for this practice even if regional data were available. Truly exogenous variables are driven by broad economy-wide forces and should not contain endogenous local fluctuations. For the quarterly agricultural regressions of Section IV it was possible to compute regional agricultural prices but only on an annual basis. Quarterly models with these prices repeated four times gave much lower $\bar{R}^2$ than the reported equations. Evidently, the time variation is more important than the regional variation.

[5] In the Northeast, state minimum wages tend to be just at the federal wage, and outside the Northeast they are usually well below the federal wage. Movements in state wage floors tend to parallel changes in the federal wage, and the main effect of state minimums is to extend coverage to industries only partly covered by federal legislation: retail trade, laundries, services, etc., industries excluded from the sample. In 1970 the federal minimum was $1.60, while the simple averages of state minimums by region were: New England, $1.60; Middle Atlantic, $1.60; East North Central, $.98; West North Central, $.65; South Atlantic, $.77; East South Central, $.18; West South Central, $.84; Mountain, $1.07; and Pacific, $1.50. New York was the only state with a minimum wage above $1.60; its minimum was raised from $1.60 to $1.85 effective July 1, 1970. Washington, D.C., and Alaska had 1970 minimums of $1.80 and $2.00, respectively, but were excluded from the sample. Source: U.S. Employment Standards Administration (1972*b*, p. 52, table 5).

[6] Many of these studies use the ratio of teenage to adult employment as the dependent variable, and so they estimate only a pure substitution effect (Welch 1974). This study estimates a combined scale and substitution effect.

TABLE 3

Wage Distribution for the Aggregation of Four Nondurable Manufacturing Industries, March 1964 (%)

Straight-Time Wage ($) Less Than	United States	Northeast	South	North Central	West
1.25	1	1	1	1	1
1.30	16	11	22	18	14
1.35	23	16	30	23	19
1.40	29	21	39	31	22
1.45	36	28	47	37	28
1.50	41	32	53	42	31
2.50	91	85	98	94	83
3.50	98	96	100	100	97
Total	100	100	100	100	100
Average hourly earnings ($)	1.76	1.90	1.59	1.72	1.92
Under $1.40 shares of total:					
Employment (k_1)	.29	.21	.39	.31	.22
Labor costs (s_1)	.22	.14	.32	.24	.15

Note.—In September 1963 the basic minimum wage was raised from $1.15 to $1.25; in February 1967 it was raised to $1.40. The distributions were computed from table 4 of U.S. Wage and Hour and Public Contracts Divisions (1965). Accounting identities and interpolations were used to get complete data for the North Central and West regions, and so the distributions for these two regions are less reliable than the other distributions.

collected in a special survey of firms by the Department of Labor. Define low-wage workers to be those within 10 percent of the minimum wage of $1.25 (10 percent was about the average percentage increase in the minimum wage in the early 1960s). Then the share of low-wage labor in total labor costs is 0.14 in the Northeast and 0.32 in the South. These cost shares indicate that, if firms held employment constant, then a 10 percent minimum-wage hike would drive up labor costs by 1.4 and 3.2 percent in these two regions. Furthermore, if firms respond to the wage hike by contracting low- and high-wage employment equiproportionately, then the average wages in the North and South would increase by 1.4 and 3.2 percent. According to the regressions in table 2, the average wages actually increase by 1.5 percent in the North and 3.2 percent in the South. (These figures are 10 times the unweighted mean elasticity for the two northern regions and three southern regions.) A similar agreement between the cost share and the wage elasticity exists in the North Central region and in the West. The data appear to say that covered-sector firms respond to a minimum-wage hike by taking steps to reduce total labor input and that both groups of workers suffer equiproportionate disemployment. Furthermore, since the substitution effect appears to be small

relative to the scale effect, the estimated disemployment elasticities are close to the true demand elasticities for low-wage labor.

IV. The Uncovered Sector

The purpose of this section is to estimate the minimum-wage effect on uncovered (agricultural) employment and wages. The reduced-form equations (2a) and (2b) give the specifications. Specifically, the logs of agricultural employment and wages are regressed on the minimum wage, the control wage, the general price level, and the shift variables appearing in (1a) and (1b). Data are available quarterly, and the operational definitions are as follows. Employment and wages, n^u and w^u, are, respectively, the logs of hired farm employment and the hourly wage exclusive of room and board. Agricultural economists consider this wage to be more reliable than the U.S. Department of Agriculture's various composite measures. In the list of independent variables, the minimum wage variable is cw, the legal minimum multiplied (before logging) by the Moore index of the proportion of nonagricultural employees covered by the legislation. Intuitively, the impact of a 10 percent minimum-wage hike on the uncovered agricultural labor market is much different if 20 versus 90 percent of the nonagricultural work force is covered. The control wage w_2 and the cyclical index z are the same variables as used in Section III. The general price level p is the log of the CPI. Finally, the uncovered-sector shift variables are the logs of the USDA's index of prices received by farmers, p^u, and the index of prices paid by farmers for production items q. The index q excludes wages.[7]

The basic sample period (before lags and leads) is 1947:I–1966:IV. Truncation at 1966 restricts the estimation to a period in which agriculture was effectively uncovered by federal legislation. On February 1, 1967, coverage was extended to employees of large farms. The Department of Labor estimated that by 1971 about 48 percent of the agricultural work force was covered by federal legislation.[8] Also, after extension of federal coverage to agriculture, several states expanded their coverage to agriculture. Prior to 1967, state agricultural minimums were considered unimportant.[9]

[7] The agricultural employment and wage series were obtained from various issues of *Farm Labor* (U.S. Department of Agriculture 1951–75). The input and output prices indexes were obtained from issues of *Agriculture Prices, Annual Summary* (U.S. Department of Agriculture 1961–72).

[8] For details see U.S. Employment Standards Administration (1972*a*, pp. 5–9).

[9] In 1966 only three states in the sample had a minimum wage applicable to agricultural workers: California, $1.30; Michigan, $1.15; and Wisconsin, $1.00. The California and Wisconsin minimums applied only to women and minors. Source: U.S. Wage and Hour and Public Contracts Divisions (1966, p. 30, table 12). Michigan and Wiscon-

Two conclusions are apparent from the regressions. First, the significance tests in table 4 confirm the well-known fact that aggregate cyclical forces play a strong role in the determination of agricultural labor supply. According to the U.S. long-run elasticities in table 5, a 10 percent increase in industrial production is associated with an 11.3 percent decrease in hired farm employment and a 2.9 percent increase in the hired farm wage. This is important because it indicates that there are indeed sizable labor flows between agriculture and the covered sector.

Second, and most important, changes in the minimum wage do affect agricultural employment and wages. The F-tests in table 4 show that as a group the distributed lag coefficients on the minimum wage are significant at 10 percent in five of 10 employment equations and in eight of 10 wage equations. Note that in table 5 the long-run U.S. employment elasticity of $-.18$ is insignificant, but the wage elasticity of $-.08$ is significant. The latter measurement indicates that for the entire United States a minimum-wage hike tends to depress the uncovered-sector wage. This finding is consistent with both Mincer's and Gardner's results.

Examination of the regional results in table 5 suggests that there are important regional differences in the welfare effects of the minimum wage. Notice that the effect on the uncovered-sector wage is significantly positive in the New England, Middle Atlantic, and West North Central regions; it is significantly negative in the South Atlantic and Pacific regions. Thus, as indicated by (3a), the wage in the uncovered sector can indeed rise or fall after a minimum-wage hike in the covered sector. Viewed in isolation, however, the uncovered-sector results are insufficient to verify the model. Indeed, if the hypothesized economic mechanisms are really at work, then the absolute disemployment elasticities from Section III should be largest in the regions where the uncovered-sector wage falls.

V. Sectoral Comparisons

Consider the inequality (3a). If the demand elasticity e^c is less than the turnover parameter δ, then after a minimum-wage hike the flow of labor out of the uncovered sector increases the wage there; if e^c is greater than δ, then the flow of labor into the uncovered sector decreases the wage. Therefore, if there is small regional variation in

sin are in the East North Central region, and their agricultural minimum wages possibly could explain the results for that region. Consider, though, the fact that California accounts for 88 percent of agricultural employment in the Pacific region, and it turns out that after a hike in the U.S. minimum wage the agricultural wage *declines* in the Pacific region.

TABLE 4

UNCOVERED-SECTOR EMPLOYMENT AND WAGE EQUATIONS: SIGNIFICANCE TESTS

	DEPENDENT VARIABLE n^u (1948:IV–1966:I)						DEPENDENT VARIABLE w^u (1948:IV–1966:I)					
	cw [0–10]	w_2 [0–5]	z [0–4]	p^u [0–5]	p [0–7]	q [0–7]	cw [0–10]	w_2 [0–5]	z [0–4]	p^u [0–5]	p [0–7]	q [0–7]
United States	***	****	***	***	*	****	****	****	****	****	****	****
New England	...	...	...	...	...	...	****	****	****	****	****	****
Middle Atlantic	...	...	*	...	...	...	****	**	****	****	**	***
East North Central	...	...	***	***	*	*	**	...	****	***	**	...
West North Central	**	****	*	***	...	***	...	...	****	****	...	****
South Atlantic	**	***	***	***	**	***	****	*	****	***	***	...
East South Central	...	*	**	...	*	...	****	...	****	***	****	**
West South Central	**	**	****	...	...	***	...	...	...	...	...	...
Mountain	*	****	**	*	...	*	***	*	...	*	**	*
Pacific	****	***	***	...	...	...	****	***	****	****	***	****

NOTE.—Test statistics are F- statistics with numerator degrees of freedom equal to the number of included lag coefficients for each variable and denominator degrees of freedom equal to 18. Lags are in brackets. Dependent variables: n^u = log of hired farm employment, by region as indicated; w^u = log of hourly wage for hired farm labor exclusive of room and board, by region as indicated. Independent variables: cw = log of coverage-weighted minimum wage; w_2 = log of average hourly earnings, all manufacturing excluding average hourly earnings in the four low-wage manufacturing industries; z = log of the FRB index of industrial production, all manufacturing; p^u = log of the USDA index of prices received by farmers; p = log of the consumer price index, all items; q = log of the USDA index of prices paid by farmers for production items.

* 15%.

** 10%.

*** 5%.

**** 1%.

TABLE 5

UNCOVERED SECTOR EMPLOYMENT AND WAGE EQUATIONS: SUMS OF DISTRIBUTED LAG COEFFICIENTS (Long-Run Effects)

	Dependent Variable n^u						Dependent Variable w^u					
	cw [0–10]	w_2 [0–5]	z [0–4]	p^u [0–5]	p [0–7]	q [0–7]	cw [0–10]	w_2 [0–5]	z [0–4]	p^u [0–5]	p [0–7]	q [0–7]
United States	−.18	3.23	−1.13	1.29	−3.60	−.60	−.08	1.28	.29	.71	−1.59	−.08
	(.13)	(1.44)	(.29)	(.85)	(2.89)	(1.29)	(.02)	(.23)	(.07)	(.14)	(.47)	(.22)
New England	−.17	5.99	−1.24	−.97	−3.60	1.57	.07	1.12	.29	.78	−.92	−.82
	(.39)	(4.40)	(.90)	(2.61)	(8.77)	(3.94)	(.03)	(.35)	(.09)	(.20)	(.71)	(.30)
Middle Atlantic	.09	.87	−.95	.14	1.33	−.13	.08	.63	.26	.15	.22	−.23
	(.19)	(2.32)	(.49)	(1.30)	(4.72)	(1.83)	(.02)	(.23)	(.07)	(.13)	(.48)	(.21)
East North Central	−.15	5.82	−1.17	3.28	−7.49	−2.80	.05	−.14	.56	−.11	2.14	−.30
	(.24)	(2.54)	(.50)	(1.48)	(5.22)	(2.13)	(.04)	(.43)	(.09)	(.24)	(.89)	(.36)
West North Central	−.27	7.58	−.85	4.83	−9.85	−3.95	.06	.73	.34	.98	.41	−1.25
	(.18)	(1.99)	(.41)	(1.14)	(4.02)	(1.63)	(.03)	(.37)	(.08)	(.20)	(.77)	(.28)
South Atlantic	.09	2.10	−1.12	1.55	−5.79	−1.12	−.10	−1.19	.79	−.53	3.55	−.17
	(.19)	(1.95)	(.42)	(1.13)	(3.98)	(1.63)	(.05)	(.56)	(.12)	(.30)	(1.21)	(.41)
East South Central	.29	−1.55	−.00	−1.07	5.07	1.78	.08	1.42	.24	.88	−2.57	−.34
	(.45)	(4.60)	(.87)	(3.41)	(8.95)	(5.26)	(.08)	(.84)	(.18)	(.55)	(1.64)	(.83)
West South Central	−.57	2.55	−2.26	1.62	1.87	−2.70	−.05	.06	.50	.19	−.91	.58
	(.31)	(4.40)	(.68)	(2.25)	(9.46)	(3.67)	(.04)	(.47)	(.11)	(.29)	(.94)	(.44)
Mountain	−.11	5.83	−.07	1.89	−9.76	−.77	−.07	.62	.13	1.00	−2.07	−.57
	(.17)	(1.73)	(.37)	(.98)	(3.60)	(1.42)	(.06)	(.70)	(.14)	(.50)	(1.38)	(.75)
Pacific	−.15	2.45	−1.14	.79	−1.48	−.33	−.12	.75	.49	.84	.13	−1.16
	(.16)	(1.70)	(.35)	(1.00)	(3.40)	(1.51)	(.03)	(.32)	(.08)	(.18)	(.66)	(.28)

NOTE.—SE in parentheses. In both equations, df = 18. See table 4 for definitions of variables. Lags are in brackets.

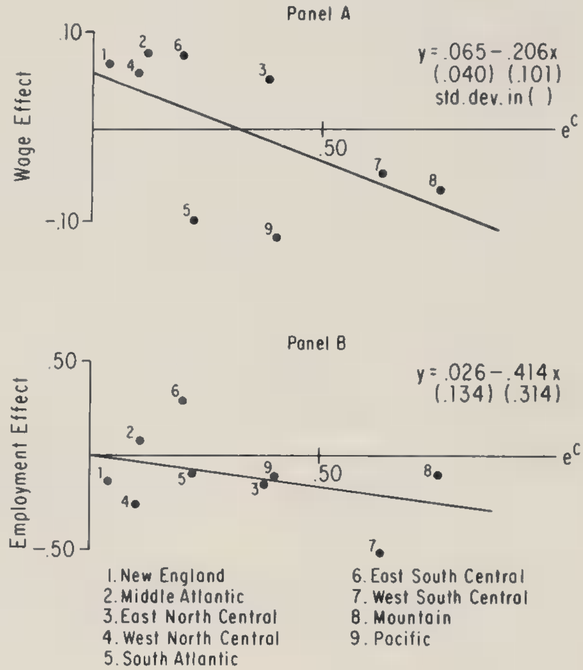


FIG. 1.—The relationship between the uncovered-sector wage and employment effects and the covered-sector disemployment elasticity.

the parameter δ, then plotting regional observations on the effect of the minimum wage on the wage in the uncovered sector against the demand elasticity e^c should give a downward sloping relationship crossing the horizontal axis at δ. By (3b), plotting the effect of the minimum wage on employment in the uncovered sector against e^c should give an upward sloping relationship also crossing the horizontal axis at δ. Panels *A* and *B* of figure 1 are the corresponding empirical diagrams. The dependent variable in panel *A* is the observed long-run elasticity of the uncovered-sector wage with respect to the minimum wage (from table 5); the dependent variable in panel *B* is the long-run uncovered-sector employment elasticity (also from table 5). The independent variable in each panel is the absolute value of the long-run covered-sector disemployment elasticity (from table 2).

As predicted by theory, panel *A* indicates an inverse relationship between the effect of the minimum wage on the uncovered-sector wage and the demand elasticity. The fitted line's horizontal intercept of 0.32 is an estimate of the critical demand elasticity (δ) that determines the sign of the wage effect in the uncovered sector. The estimate $\delta = 0.32$ says that in the steady state at least 32 percent of covered-sector employees compete with the unemployed for employment. In contrast, panel *B* suggests no relationship between the

effect of the minimum wage on uncovered employment and the demand elasticity. The data points are essentially random deviations about the horizontal axis. Note that the F-tests in table 4 indicate that the minimum wage affects uncovered employment. The sign and size of the long-run effect, however, are measured very imprecisely.

The skill-group aggregation in the uncovered sector data can account for the apparent conflict between theory and panel B of figure 1. When the low-skill wage changes in the uncovered sector, there are scale and substitution effects. As noted before and as proved in Appendix A, the larger is the substitution effect relative to the scale effect, the smaller is the observed (absolute) employment elasticity relative to the true proportionate change in low-wage labor. Evidently, the elasticity of substitution between low- and high-skill labor in agriculture is large enough to make this skill-group aggregation suppress the uncovered employment effect to an undetectable level. Notice that this is the reverse of the situation in the covered sector. As was noted in Section III, in the covered sector the scale effect dominates the substitution effect; in contrast, in the uncovered sector the substitution effect dominates the scale effect.

VI. Summary and Conclusion

This paper examines the covariation over regions in the effects of the minimum wage on wages and employment in the covered and uncovered sectors. Its major empirical finding is that the pattern of results is consistent with theory. Specifically, after a minimum-wage hike the uncovered-sector wage increases in the five regions with relatively small covered-sector demand elasticities: New England, Middle Atlantic, East North Central, West North Central, and East South Central; it decreases in the four regions with the relatively large covered-sector demand elasticities: South Atlantic, West South Central, Mountain, and Pacific. In addition, Mincer's and Gardner's earlier work suggesting that the uncovered-sector wage decreases in the entire United States is consistent with these findings, for their results are dominated by the four large regions (in terms of low-wage employment) in which the uncovered-sector wage falls. This paper's results have some bearing on the issue of welfare effects of the minimum wage. It seems fair to conclude that the minimum wage increases the well-being of all low-wage workers in the New England, Middle Atlantic, East North Central, West North Central, and East South Central regions. In the remaining regions, however, the minimum wage increases the well-being of only those workers having covered-sector employment.

Appendix A

The purpose of this Appendix is to characterize the effects of using total employment and the average wage as dependent variables in the regressions. Suppose low-skill class 1 labor (N_1) and high-skill class 2 labor (N_2) are imperfect substitutes in a production function with constant returns to scale. Let the class 1 wage (W_1) increase. Then

$$\tilde{N}_1 = (\text{scale effect} - s_2\sigma_{12})\tilde{W}_1,$$

$$\tilde{N}_2 = (\text{scale effect} + s_1\sigma_{12})\tilde{W}_1,$$

where $\tilde{}$ denotes proportionate change, s_1 and s_2 are the factors' shares in expenditures, and $\sigma_{12} > 0$ is the absolute Hicks-Allen partial elasticity of substitution. Let k_1 and k_2 denote the factors' shares in total employment; then the proportionate change in total employment is

$$\begin{aligned}\tilde{N} &= k_1\tilde{N}_1 + k_2\tilde{N}_2 \\ &= [\text{scale effect} + (k_2 - s_2)\sigma_{12}]\tilde{W}_1,\end{aligned}$$

which is less negative than $\tilde{N}_1$. The change in the average wage is

$$\begin{aligned}\tilde{\overline{W}} &= (k_1W_1 + k_2W_2) \\ &= s_1\tilde{W}_1 + (s_1 - k_1)(\tilde{N}_1 - \tilde{N}_2) \\ &= [s_1 - (s_1 - k_1)\sigma_{12}]\tilde{W}_1,\end{aligned}$$

which exceeds $s_1\tilde{W}_1$ because $s_1 < k_1$ (class 1 labor is the low-wage factor). Clearly the characterization applies equally well to either sector. (The above arguments were found by an anonymous referee.)

Appendix B

This Appendix discusses the technical econometric issues associated with the empirical work.

Specification

There are three important points concerning the covered-sector equations. First, adjustment costs (hiring and separation costs) induce a dependence of factor demand on lagged and expected future factor prices (Sargent 1978). This motivates using distributed lag methods and including a lead value of the minimum wage, for its movements are known well in advance. Second, minor distortions due to temporal aggregation (the unavoidable use of annual instead of monthly data) are apparently the cause of a significant lead coefficient on the control wage. This was verified by using Geweke's (1978) results. The effects of the temporal aggregation are minor because the explanatory variables are annual averages and the long-run effects are therefore estimated consistently (see Geweke's theorem 2). Third, the temporal aggregation and the forecasting of factor prices are good reasons to expect the homogeneity of degree zero of (1a) in nominal variables to break down in practice. Consequently, w and w_2 are entered separately in (1a) and the homogeneity constraint tested (and rejected). The coefficients have a "real" interpretation, for the entire history of each nominal variable is held constant.

The same comments apply to the uncovered sector regressions, though

TABLE B1

Exogeneity Tests

	Covered Sector								Uncovered Sector[a]											
	Dependant Variable n^c				Dependent Variable w^c				Dependent Variable n^u						Dependent Variable w^u					
	w [−2]	w_2 [−1]	w_2 [−2]	z [−1]	w [−2]	w_2 [−1]	w_2 [−2]	z [−1]	cw	w_2	z	p^u	p	q	cw	w_2	z	p^u	p	q
United States	...	****	...	...	...	***	...	...	...	...	...	...	...	...	...	...	...	...	...	...
New England	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Middle Atlantic	...	...	...	...	...	**	...	...	...	...	...	...	***	**	...	...	**	...	...	...
East North Central	...	*	...	...	...	...	...	...	...	...	...	...	...	**	***	****	...	...	...	...
West North Central	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
South Atlantic	...	****	*	...	...	***	...	*	...	...	...	...	...	...	...	...	...	***	...	...
East South Central	...	...	...	...	...	***	...	...	...	...	...	...	...	...	...	...	...	...	...	...
West South Central	***	****	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Mountain	***	****	...	...	N.A.	N.A.	N.A.	N.A.	...	**	...	...	...	...	*	...	...	...	...	...
Pacific	...	****	...	...	...	...	...	...	...	...	**	...	...	...	...	...	**	***	...	...

Note.—Tests are for indicated lead coefficients equal to zero. The statistics are $t(17)$ in the n^c equation, $t(8)$ in the w^c equation, and F (3,18) in both the n^u and w^u equations. Leads are in brackets.

[a] All leads in the n^u and w^u equations are −1 to −3.

* 15%.

** 10%.

*** 5%.

**** 1%.

with quarterly data the temporal aggregation problems are absent. To purge the series of seasonal effects, each regression included seasonal dummies and the products of seasonals with trend.

Serial Correlation

In the annual covered-sector regressions a first-order Cochrane-Orcutt correction was used; in the quarterly uncovered-sector regressions ad hoc prefilters were used together with fifth-order Cochrane-Orcutt.

Exogeneity

Table B1 reports the outcome of batteries of Sims's (1972, 1977) exogeneity test. The test consists of the appropriate F- or t-test on lead coefficients. The paucity of rejections indicates no serious problems with a lack of exogeneity in either set of results. Since lead values of w and w_2 enter the covered-sector equations for other reasons, the test's power is obviously low in that set of results.

References

Gardner, Bruce. "Minimum Wages and the Farm Labor Market." *American J. Agricultural Econ.* 54 (August 1972): 473–76.
Geweke, John F. "Temporal Aggregation in the Multiple Regression Model." *Econometrica* 46 (May 1978): 643–61.
Gramlich, Edward M. "Impact of Minimum Wages on Other Wages, Employment, and Family Incomes." *Brookings Papers Econ. Activity*, no. 2 (1976), pp. 409–51.
Mincer, Jacob. "Unemployment Effects of Minimum Wages." *J.P.E.* 84, no. 4, pt. 2 (August 1976): S87–S104.
Sargent, Thomas J. "Estimation of Dynamic Labor Demand Schedules under Rational Expectations." *J.P.E.* 86 (December 1978): 1009–44.
Silberman, Jonathan I., and Durden, Garey C. "Determining Legislative Preferences on the Minimum Wage: An Economic Approach." *J.P.E.* 84 (April 1976): 317–29.
Sims, Christopher A. "Money, Income, and Causality." *A.E.R.* 62 (September 1972): 540–52.
———. "Exogeneity and Causal Orderings in Macroeconomic Models." In *New Methods in Business Cycle Research: Proceedings from a Conference*, edited by Christopher A. Sims. Minneapolis: Federal Reserve Bank of Minneapolis, 1977.
U.S. Bureau of Labor Statistics. *Employment and Earnings, United States, 1909–75.* Statistical Bulletin no. 1312-10. Washington: Government Printing Office, 1976.
———. *Employment and Earnings.* Washington: Government Printing Office, 1976–77, various issues.
———. *Employment and Earnings, State and Areas, 1939–75.* Statistical Bulletin no. 1370-12. Washington: Government Printing Office, 1977.
U.S. Department of Agriculture, Statistical Reporting Service. *Agricultural Prices, Annual Summary.* Washington: Government Printing Office, 1961–72, various issues.
———. Statistical Reporting Service, Agricultural Marketing Service, Bureau

of Agricultural Economics. *Farm Labor.* Washington: Government Printing Office, 1951–75, various issues.

U.S. Department of Commerce. *Business Statistics.* 1975 ed. Washington: Government Printing Office, 1976.

———. *Survey of Current Business.* Washington: Government Printing Office, 1976–77, various issues.

U.S. Employment Standards Administration. *Hired Farm Workers.* Washington: Government Printing Office, 1972. (*a*)

———. *Wages and Hours of Nonsupervisory Employees in All Private Nonfarm Industries by Coverage Status under the Fair Labor Standards Act.* Washington: Government Printing Office, 1972. (*b*)

U.S. Wage and Hour and Public Contracts Divisions. *Manufacturing Industries: A Study to Evaluate the Minimum Wage and Maximum Hours Standards of the Fair Labor Standards Act.* Washington: Government Printing Office, 1965.

———. *Hired Farm Labor: A Study to Evaluate the Minimum Wage and Maximum Hours Standards of the Fair Labor Standards Act.* Washington: Government Printing Office, 1966.

Welch, Finis. "Minimum Wage Legislation in the United States." *Econ. Inquiry* 12 (September 1974): 285–318.

Taxes, Inflation, and the Durability of Capital

Andrew B. Abel

Harvard University

The choice of capital durability is affected by the rate of inflation because nominal depreciation deductions are based on historical cost rather than replacement cost. This paper analyzes a model with competitive firms and intertemporally optimizing consumers and demonstrates that an increase in the rate of inflation will lead to an increase in the durability of capital, if the ratio of the nominal discount rate to the rate of depreciation exceeds a critical value. However, if this ratio is less than the critical value, this effect is reversed. The effect of costs of adjustment on the critical value is analyzed.

I. Introduction

Since depreciation deductions for tax purposes are based on historical cost rather than replacement cost, inflation can affect the investment and capital durability decisions of firms. This problem was recently analyzed by Auerbach (1979), using a general equilibrium model in which investment and durability are determined by optimizing firms and consumption is determined according to a proportional savings function. To analyze the effect of inflation on the choice of asset durability, Auerbach simulates his model, using a Cobb-Douglas technology with reasonable parameter values.[1] His simulation results indicate that higher inflation leads firms to choose more durable capital, although we are left with a caveat about the robustness of the results.

This paper was written while I was an assistant professor at the University of Chicago Department of Economics. I thank Alan Auerbach, Dennis Carlton, Stanley Fischer, Robert LaLonde, Frederic Mishkin, and Michael Mussa for comments on earlier drafts of this paper.

[1] Auerbach also assumes values for the share of government expenditure in net output, the fraction of disposable income saved, and the corporate tax rate.

[*Journal of Political Economy,* 1981, vol. 89, no. 3]

By modifying Auerbach's model to make consumption decisions based on intertemporal utility maximization rather than on a proportional saving function, we can derive analytically the effect of inflation on durability without having to specify the form of the production function. The direction of the effect of inflation on durability is determined by whether the ratio of the nominal discount rate to the depreciation rate exceeds a critical value which depends solely on the corporate tax rate and the rate of the investment tax credit. If the ratio of the nominal discount rate to the depreciation rate exceeds the critical value—as is the case in Auerbach's simulations[2]—then an increase in the rate of inflation leads to an increase in durability. However, if the ratio of the nominal discount rate to the depreciation rate is less than the critical value, then an increase in the rate of inflation leads to less durable capital. This reversal of Auerbach's finding is more likely to apply to durable equipment rather than to structures since equipment has a higher depreciation rate than structures.[3]

A second modification of Auerbach's model is the incorporation of adjustment costs into the model. The presence of convex costs of adjustment removes the indeterminacy of the individual firm's rate of investment in the Auerbach model. It should be noted that one complication of introducing adjustment costs is that the critical value of the ratio of the nominal discount rate to the depreciation rate is no longer a function of only tax parameters. The critical value of this ratio will, in general, depend on technology, although we can place bounds on this critical value. Using these bounds on the critical value of the ratio of the nominal discount rate to the depreciation rate, we can still derive fairly strong results about the effect of inflation on durability based only on this ratio and on the tax parameters. In general, if the nominal discount rate exceeds the rate of depreciation, an increase in the rate of inflation reduces the rate of depreciation. If the nominal discount rate is "enough" less than the depreciation rate, where enough depends only on tax parameters, an increase in the rate of inflation increases the rate of depreciation. In the region between the upper and lower bounds on the critical value, the critical value depends on technology. Note, however, that this complication arises from the introduction of adjustment costs. In the absence of adjustment costs, the critical value of the ratio of the nominal discount rate to the depreciation rate depends only on tax parameters.

[2] For a more precise discussion of Auerbach's simulation results, see Section III.

[3] This statement implicitly assumes that the investment tax credit rate is the same for structures and equipment, so that the critical value is the same for both types of capital. However, under the current U.S. tax laws, the investment tax credit is applicable to equipment but not to structures. Therefore, as will become clear in Section III, the critical value is smaller for equipment than for structures. This effect may weaken or reverse the statement in the text.

In Section II we develop a model of a competitive firm and analyze the investment and durability decisions. The model of the firm is incorporated into a general equilibrium model in Section III, where we analyze the effect of changes in the rate of inflation. Section IV presents concluding remarks.

II. The Model of the Firm

Since the model of the firm in this paper differs from Auerbach's model only by the inclusion of adjustment costs and the investment tax credit, the description of the model will be brief; further description may be found in Auerbach.

We assume that the firm is a price taker in its output market and in the labor market. The price of output p and the nominal wage rate w both inflate at rate π, so that the real wage is constant over time. Let δ_t be the (constant) rate of exponential decay of capital goods installed at time t. Ex ante, δ_t is a decision variable of the firm. We assume that the instantaneous flow of capital services from a unit of capital is an increasing concave function of δ, $A(\delta)$, with $A' > 0$, $A'' < 0$.

Investment in physical capital consists of purchasing a unit of the homogeneous output and installing it in place at some cost which is an increasing convex function of the rate of investment. Let I_t be nominal investment at time t, so that I_t/p_t is the number of units of output bought and installed as physical capital. We assume that the real cost of installation is an increasing convex function of real investment, so that the nominal cost of installation at time t is $p_t c(I_t/p_t)$, with $c(0) = c'(0) = 0$, and $c' > 0$, $c'' > 0$ for $I_t/p_t > 0$.[4] We assume that these installation costs can be expensed for tax purposes. Note that the assumption that the purchase of uninstalled capital goods is amortized over time while the installation cost is immediately expensed is similar to Auerbach's case in which a fraction e of the cost of investment is expensed.[5]

Capital services at time s are

$$KS_s = \int_{-\infty}^{s} A(\delta_t)(I_t/p_t)e^{-\delta_t(s-t)}dt \tag{1}$$

since $(I_t/p_t)e^{-\delta_t(s-t)}$ is the amount of capital installed at time t which remains at time s and $A(\delta_t)$ is the flow of capital services per unit of this

[4] Adjustment costs have been incorporated into models of investment by Eisner and Strotz (1963), Lucas (1967), Gould (1968), and Treadway (1969). Abel and Blanchard (1980) distinguish the cost of purchasing uninstalled capital from the cost of installing capital, as is done in the text.

[5] There is a slight difference, however, in that the fraction of the total cost which is expensed in this model is endogenous, whereas the fraction e is exogenous in Auerbach's model.

capital. Gross output at time s is given by $Y_s^G = H(KS_s, L_s)$, where H is first-degree homogeneous in KS and L. We also assume $H_{KS}, H_L, H_{KS,L} > 0$ and $H_{KS,KS}, H_{L,L} < 0$.

Taxable corporate profits, which are revenues less wages, installation costs, and depreciation allowances, are taxed at rate τ, $0 < \tau < 1$. Let $D(x, \delta)$ be the depreciation deduction per dollar of purchase cost of capital of age x which decays at rate δ, and let k $(0 \leq k < 1)$ be the investment tax credit rate applied to the purchase of uninstalled capital.

The objective of the firm is to maximize

$$V = \int_0^\infty \Big\{(1-\tau)[p_s Y_s^G - w_s L_s - p_s c(I_s/p_s)] - (1-k)I_s + \tau \int_{-\infty}^{s} I_v D(s-v, \delta_v)dv\Big\} e^{-rs}ds \tag{2}$$

with respect to L_s, δ_s, and I_s, where r is the rate at which equity holders discount after-tax nominal flows.[6] Note that (2) is the same as Auerbach's equation 3 except for the inclusion of installation costs and the investment tax credit. For simplicity of notation, let

$$z(\delta) = \int_0^\infty D(x, \delta)e^{-rx}dx \tag{3}$$

be the present value of depreciation deductions, and let $\rho = r - \pi$ be the after-tax real discount rate. Differentiating V with respect to δ_t, I_t, and L_t and setting these derivatives equal to zero, we obtain:[7]

$$(1-\tau)\int_t^\infty [A'(\delta_t) - (s-t)A(\delta_t)]H_{KS}(KS_s, L_s)e^{-(\rho+\delta_t)(s-t)}ds = \frac{\partial}{\partial \delta_t}[1 - k - \tau z(\delta_t)], \tag{4a}$$

$$(1-\tau)c'(I_t/p_t) = (1-\tau)A(\delta_t)\int_t^\infty H_{KS}(KS_s, L_s)e^{-(\rho+\delta_t)(s-t)}ds - [1 - k - \tau z(\delta_t)], \tag{4b}$$

$$H_L(KS_t, L_t) = w_t/p_t. \tag{4c}$$

These equations may be solved recursively.[8] At any point in time, t, KS_t is fixed and the firm chooses L_t to satisfy (4c). Note that, since the

[6] We assume that firms are financed entirely by equity, so that interest payments are not included in the expression for cash flow.

[7] The first-order conditions are identical to Auerbach's eq. 4, except that $(1 - \tau)c'(I_t/p_t)$ in (4b) and k in (4a) and (4b) are zero in Auerbach's framework.

[8] If the installation costs were amortized over time rather than immediately expensed, then the system would be only block recursive. Eq. (4c) could be solved first, but the analogues of (4a) and (4b) would have to be solved simultaneously.

real wage, w_t/p_t, is assumed to be constant over time, $H_L(KS_t,L_t)$ is constant over time. Since $H(KS_t,L_t)$ is linearly homogeneous, the constancy of $H_L(KS_t,L_t)$ implies that $H_{KS}(KS_t,L_t)$ is constant over time. Let ϕ be the constant value of $H_{KS}(KS_t,L_t)$ and define $\lambda(\delta;\phi)$ as

$$\lambda(\delta_t;\phi) = \frac{(1-\tau)A(\delta_t)\phi}{\rho+\delta_t} + k + \tau z(\delta_t) - 1. \tag{5}$$

Note that $[(1-\tau)A(\delta_t)\phi]/(\rho+\delta_t)$ is the present value of the stream of after-tax marginal products accruing to the undepreciated portion of a unit of capital installed at time t. In addition, by installing a unit of capital at time t, the firm reduces the present value of its taxes by $k + \tau z(\delta_t)$. Therefore, $\lambda(\delta;\phi)$ is the excess of the present value of the after-tax cash flow accruing to the marginal unit of installed capital over the cost of an uninstalled unit of capital.

Note that the first-order conditions given in (4a) and (4b) may be expressed as[9]

$$\lambda'(\delta_t;\phi) = \frac{(1-\tau)\phi}{\rho+\delta_t}\left[A'(\delta_t) - \frac{A(\delta_t)}{\rho+\delta_t}\right] + \tau\frac{\partial z(\delta_t)}{\partial\delta_t} = 0, \tag{6a}$$

$$(1-\tau)c'(I_t/p_t) = \lambda(\delta_t;\phi). \tag{6b}$$

To summarize, the firm may use the following sequential decision process. Given a real wage rate, the firm chooses a KS/L ratio to equate H_L with the exogenous real wage. This KS/L ratio uniquely determines ϕ, the marginal product of capital services. Given ϕ, the excess of the shadow value of installed capital over the net cost of uninstalled capital is a function, $\lambda(\delta_t;\phi)$, of δ_t. According to (6a), the firm chooses δ_t to maximize $\lambda(\delta_t;\phi)$.[10] Given the resulting maximized value of $\lambda(\delta_t;\phi)$, the firm will choose the real rate of investment to equate the marginal installation cost with $\lambda(\delta_t;\phi)$, as in (6b). This solution procedure applies at all points in time, even when the capital stock has not yet reached its steady-state value. Since the real wage rate is assumed constant, the marginal product of capital services, ϕ, is also constant. Therefore, the λ-maximizing value of δ is constant over time, and the

[9] If we impose the constraint that gross investment is nonnegative, the first-order conditions with respect to I are: (*a*) $(1-\tau)c'(I_t/p_t) \geq \lambda(\delta;\phi)$; and (*b*) $(I_t/p_t)[(1-\tau)c'(I_t/p_t) - \lambda(\delta_t;\phi)] = 0$. If $I_t/p_t > 0$, then (*b*) implies eq. (6b).

[10] If $z = \delta/(\rho+\delta)$, as will be the case later in this note, then $\lambda(\delta;\phi) = (k-1) + [(1-\tau)A(\delta)\phi/(\rho+\delta)] + [\tau\delta/(\rho+\delta)]$, $\lambda'(\delta;\phi) = [1/(\rho+\delta)^2]\{\phi(1-\tau)[A'(\delta)(\rho+\delta) - A(\delta)] + \tau\rho\}$, and $\lambda''(\delta;\phi) = [-2/(\rho+\delta)^3]\{(1-\tau)\phi[A'(\delta)(\rho+\delta) - A(\delta)] + \tau\rho\} + [(1-\tau)\phi A''(\delta)/(\rho+\delta)]$. Therefore, when $\lambda'(\delta;\phi)$ equals zero, the term in braces in the expression for $\lambda''(\delta;\phi)$ is zero and $\lambda''(\delta;\phi) = [(1-\tau)\phi A''(\delta)/(\rho+\delta)] < 0$. Hence, setting $\lambda'(\delta;\phi)$ equal to zero maximizes $\lambda(\delta;\phi)$ with respect to δ. Note that later, when we examine the case in which $z = [\delta/(\rho+\pi+\delta)]$, a similar line of reasoning will establish that $\lambda''(\delta;\phi) < 0$ when $\lambda'(\delta;\phi) = 0$, as long as π is not too large.

real rate of investment is also constant.[11] Given the constant value of real investment, say $(I/p)^*$, the steady-state capital stock is $(1/\delta)(I/p)^*$, and the steady-state flow of capital services is $[A(\delta)/\delta](I/p)^*$.

III. A General Equilibrium Model of Investment and Durability

In this section, we examine the effects of changes in the rate of inflation within the context of a general equilibrium model. Rather than posit a proportional savings function, as in Auerbach, we assume that consumption is determined as the outcome of infinite-horizon intertemporal utility maximization by identical households. As noted by Auerbach, the after-tax discount rate ρ will be constant in such a model.

To keep the analysis simple but without any essential loss of generality, assume that the number of households is constant, and each household owns one firm and inelastically supplies $\bar{L}$ units of labor to its firm. Each household selects c_t and m_t to maximize

$$\int_0^\infty U(c_t, m_t)e^{-\rho t}dt \tag{7}$$

subject to $c_t = y_t + (w_t/p_t)\bar{L} - (\dot{M}_t/p_t)$, where c_t is real consumption at time t, m_t is real balances at time t, $\dot{M}_t$ is the growth in nominal money balances, and $y_t \equiv (1-\tau)[H(KS_t,\bar{L}) - (w_t/p_t)\bar{L} - c(I_t/p_t)] - (1-k)I_t + \tau\int_{-\infty}^{t} I_s/p_t e^{-\pi(t-s)}D(t-s,\delta_s)ds$ is the after-tax real cash flow of the firm. The maximization problem is based on the assumption that each household receives the net cash flow of the firm in addition to wage income and uses its total income either to consume or to add to its money balances. We also assume that the government uses tax revenue and the issuance of new money to purchase current output—which does not affect the utility function $U(c_t, m_t)$.[12]

Solving the maximization problem (7), we can show that, in the steady state, equations (6a) and (6b) continue to hold. However, in this general equilibrium model with labor supply constant, ϕ, the marginal product of capital services, is a decreasing function of the steady-state level of capital services $[A(\delta)/\delta](I/p)$. Therefore, for the steady-state values of the variables of interest, ϕ, δ, and I/p, we have the following system of equations:

[11] Gould (1968) showed that, for a price-taking firm with constant returns to scale, the rate of investment is constant if the real wage rate is constant.

[12] Abel and Blanchard (1980) present a more complete discussion of a general equilibrium model with taxes and adjustment costs. However, in that model the depreciation rate is fixed exogenously, and money balances do not appear.

$$\phi = g\left[\frac{A(\delta)}{\delta} I/p\right]; \quad g' < 0, \tag{8a}$$

$$\lambda'(\delta, \phi) = 0, \tag{8b}$$

$$\lambda(\delta, \phi) - (1 - \tau)c'(I/p) = 0. \tag{8c}$$

Note that the system (8a)–(8c) could also represent the behavior of a single profit-maximizing firm with constant returns to scale with either some monopoly power in its output market or some monopsony power in the labor market.

To examine the effect of a change in the rate of inflation, we linearize the system (8a)–(8c) around its solution to obtain

$$\begin{bmatrix} 1 & -g' \cdot \dfrac{I/p}{\delta}\left[A'(\delta) - \dfrac{A(\delta)}{\delta}\right] & -g' \cdot \dfrac{A(\delta)}{\delta} \\ \lambda'_\phi & \lambda'_\delta & 0 \\ \lambda_\phi & 0 & -(1-\tau)c'' \end{bmatrix} \begin{bmatrix} d\phi \\ d\delta \\ d(I/p) \end{bmatrix} = \begin{bmatrix} 0 \\ -\lambda'_\pi \\ -\lambda_\pi \end{bmatrix} d\pi, \tag{9}$$

where $\lambda_\phi = [(1-\tau)A(\delta)/(\rho + \delta)] > 0$, and $\lambda'_\phi = (1-\tau)/(\rho+\delta)\{A'(\delta) - [A(\delta)/(\rho+\delta)]\} < 0$, and $\lambda'_\delta < 0$ because of the second-order condition for the optimal δ.[13] Let Δ be the determinant of the square matrix in (9), so that

$$\Delta = -\left\{\lambda'_\delta + \lambda'_\phi g' \cdot \frac{I/p}{\delta}\left[A'(\delta) - \frac{A(\delta)}{\delta}\right]\right\}(1-\tau)c'' + g' \cdot \frac{A(\delta)}{\delta}\lambda_\phi\lambda'_\delta > 0. \tag{10}$$

Note that $\Delta > 0$ regardless of whether c'' is positive or zero.

Given the linearized system (9), Cramer's rule can be used to determine the effect on δ of changes in the rate of inflation π,

$$\frac{d\delta}{d\pi} = \frac{1}{\Delta}\left[(1-\tau)c''\lambda'_\pi - g' \cdot \frac{A(\delta)}{\delta}B_\pi\right], \tag{11}$$

where $B_\pi = -\lambda'_\phi\lambda_\pi + \lambda_\phi\lambda'_\pi$. In analyzing (11), we will also consider the following two special cases: (*a*) the absence of adjustment costs, so that $c'' = 0$; and (*b*) the polar opposite case of extremely convex adjustment costs, so that $c'' = \infty$. In the case without adjustment costs, which corresponds most closely to Auerbach's model, it can be shown that

$$\frac{d\delta}{d\pi} = \frac{-1}{\Delta}g' \cdot \frac{A(\delta)}{\delta}B_\pi \quad (c'' = 0), \tag{12}$$

where the coefficient on B_π is positive. In the polar opposite case, it can be shown that

[13] Note that the element in the third row, second column of the square matrix in (9) is λ_δ, which is zero according to (8b).

$$\operatorname{sign}\frac{d\delta}{d\pi} = \operatorname{sign}\lambda'_\pi \quad (c'' = \infty). \tag{13}$$

In order to analyze the expressions for $d\delta/d\pi$, we must explicitly model the effect of inflation on the shadow price of installed capital. Following Auerbach, we assume that depreciation deductions reflect physical decay of capital but are based on historical cost, so that[14]

$$z(\delta) = \frac{\delta}{\rho + \pi + \delta}. \tag{14}$$

Using this specification of depreciation deductions, we can easily show that

$$\lambda = \frac{(1-\tau)A(\delta)\phi}{\rho + \delta} + k + \frac{\tau\delta}{\rho + \pi + \delta} - 1, \tag{15a}$$

$$\lambda' = \frac{(1-\tau)\phi}{\rho + \delta}\left[A'(\delta) - \frac{A(\delta)}{\rho + \delta}\right] + \frac{\tau(\rho + \pi)}{(\rho + \pi + \delta)^2}. \tag{15b}$$

Now we must calculate λ_π, λ'_π, and B_π. Differentiating λ with respect to π, we obtain

$$\lambda_\pi = \frac{-\tau\delta}{(\rho + \pi + \delta)^2} < 0. \tag{16}$$

Note that λ_π is negative because an increase in the rate of inflation reduces the present value of real depreciation deductions and hence reduces the shadow price of installed capital. Next, we differentiate λ' with respect to π to obtain

$$\lambda'_\pi = \frac{\tau}{(\rho + \pi + \delta)^3}(\delta - \rho - \pi) \gtreqless 0 \quad \text{as } \delta \gtreqless \rho + \pi. \tag{17}$$

Thus the sign of λ'_π depends on whether the rate of depreciation exceeds the nominal discount rate. As shown in figure 1, λ'_π is negative in region *I* and is positive in regions *II* and *III*.

Recall that $B_\pi = -\lambda'_\phi\lambda_\pi + \lambda_\phi\lambda'_\pi$, that λ_ϕ is positive, and that λ_π and λ'_ϕ are negative. Thus in region *I*, where $\delta < \rho + \pi$ and λ'_π is negative, B_π is negative. To determine the sign of B_π when $\delta > \rho + \pi$ requires more calculation. Using the expressions for λ_ϕ, λ'_ϕ, λ_π, and λ'_π, we can show that

$$B_\pi = \frac{1-\tau}{\rho + \delta}\,\frac{\tau A(\delta)}{(\rho + \pi + \delta)^2}\left\{\delta\left[\frac{A'(\delta)}{A(\delta)} - \frac{1}{\rho + \delta}\right] + \frac{\delta - \rho - \pi}{\rho + \pi + \delta}\right\}. \tag{18}$$

[14] If depreciation deductions are based on replacement cost rather than historical cost, then $z(\delta) = \delta/(\rho + \delta)$. In this case, changes in the rate of inflation have no effect on the firm's decision problem.

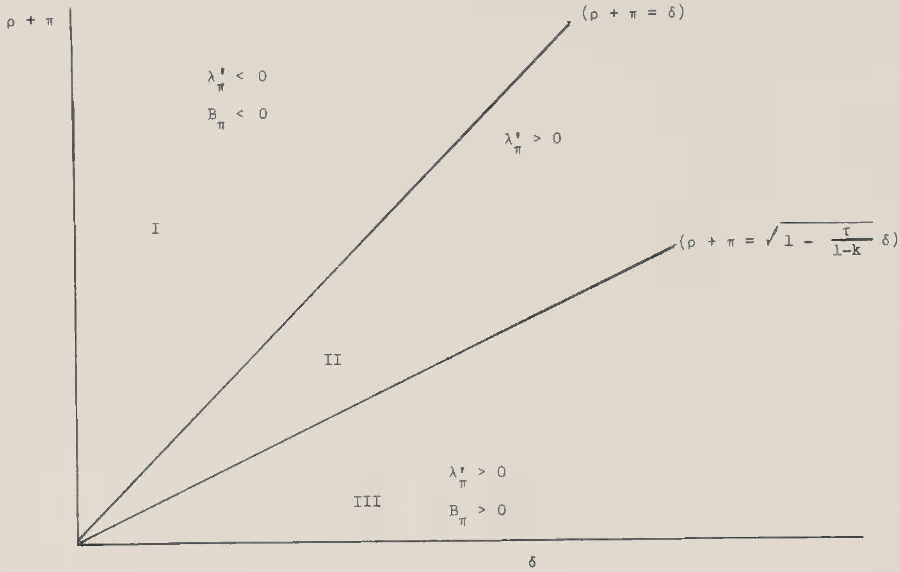


FIG. 1

It will now be useful to define q, the net cost of installed capital, ignoring the depreciation allowance, as

$$q = 1 - k + (1 - \tau)c'. \tag{19}$$

Using this expression for q, we obtain[15]

$$B_\pi = \frac{1 - \tau}{\rho + \delta} \frac{\tau A(\delta)}{(\rho + \pi + \delta)^3} \frac{q}{q(\rho + \pi + \delta) - \tau\delta} \left[\left(1 - \frac{\tau}{q}\right) \delta^2 - (\rho + \pi)^2\right]. \tag{20}$$

Note that B_π has the same sign as the term in brackets, so that

$$B_\pi \gtreqless 0 \text{ as } \frac{\rho + \pi}{\delta} \lesseqgtr \sqrt{1 - \frac{\tau}{q}}. \tag{21}$$

Thus the sign of B_π depends on whether $(\rho + \pi)/\delta$ exceeds the critical value $\sqrt{1 - (\tau/q)}$. Note that the critical value $\sqrt{1 - (\tau/q)}$ is less than or

[15] Rewrite eqq. (6a) and (6b) as

$$\frac{(1 - \tau)\phi}{\rho + \delta} \left[A'(\delta) - \frac{A(\delta)}{\rho + \delta}\right] = \frac{-\tau(\rho + \pi)}{(\rho + \pi + \delta)^2}, \tag{6a'}$$

and

$$\frac{(1 - \tau)\phi}{\rho + \delta} A(\delta) = q - \frac{\tau\delta}{\rho + \pi + \delta}. \tag{6b'}$$

Dividing (6a′) by (6b′), we obtain

$$\frac{A'(\delta)}{A(\delta)} - \frac{1}{\rho + \delta} = \frac{-\tau(\rho + \pi)}{q(\rho + \pi + \delta)^2 - \tau\delta(\rho + \pi + \delta)},$$

which may be substituted into (18).

TABLE 1

THE EFFECT OF INFLATION ON DURABILITY

	Sign of $\frac{d\delta}{d\pi}$		
Region	*I* $\frac{\rho+\pi}{\delta} > 1$	*II* $\sqrt{1-\frac{\tau}{1-k}} < \frac{\rho+\pi}{\delta} < 1$	*III* $\frac{\rho+\pi}{\delta} < \sqrt{1-\frac{\tau}{1-k}}$
General $c'' \geqslant 0$	−	?	+
No adjustment costs	−	−	+
$c'' = \infty$	−	+	+

equal to one, so that if $\rho + \pi > \delta$, as in region *I*, then $(\rho + \pi)/\delta$ exceeds the critical value, and B_π is negative. Also note that, since $q \geqslant 1 - k$, with equality holding in the absence of adjustment costs, the critical value must be greater than or equal to $\sqrt{1 - [\tau/(1 - k)]}$. Therefore, if $(\rho + \pi)/\delta < \sqrt{1 - [\tau/(1 - k)]}$ as in region *III*, B_π is positive.

Recall from (11) that, in the general case, if λ'_π and B_π have the same sign, then $d\delta/d\pi$ also has that sign. Thus, as summarized in table 1, $d\delta/d\pi$ is negative in region *I* and is positive in region *III*. In region *II*, λ'_π is positive, but the sign of B_π is indeterminate without further assumptions. If we assume that $c'' = \infty$, then the sign of $d\delta/d\pi$ is the same as the sign of λ'_π and hence is positive. However, if we assume that there are no adjustment costs, then the critical value is $\sqrt{1 - [\tau/(1 - k)]}$, and B_π is negative. Hence, in the absence of adjustment costs, $d\delta/d\pi$ is negative in region *II*.

To summarize, if the nominal discount rate exceeds the rate of depreciation, then an increase in the rate of inflation reduces the optimal δ and hence leads to more durable capital. On the other hand, if the nominal discount rate is less than $\sqrt{1 - [\tau/(1 - k)]}$ times the depreciation rate, an increase in the rate of inflation leads to less durable capital. In the case in which there are no adjustment costs and the investment tax credit is zero (as in Auerbach), an increase in the rate of inflation increases the durability of capital if $\rho + \pi > \sqrt{1 - \tau}\,\delta$ but reduces the durability of capital if $\rho + \pi < \sqrt{1 - \tau}\,\delta$.

It is interesting to compare the results for the case with zero adjustment cost and zero investment tax credit with the simulation results in Auerbach. Note that these two models are not identical in that they treat consumption decisions and money-holding decisions differently. The essential difference is that, in the optimizing model of this paper, ρ is held constant, whereas ρ is a variable in the Auerbach model. In Auerbach's table 1, the nominal discount rate exceeds the depreciation rate in every row except the first. Consistent

with the results derived above, an increase in the rate of inflation increases the durability of capital in Auerbach's simulations.

In the first row of Auerbach's table 1, the nominal discount rate falls short of $\sqrt{1-\tau}\,\delta$, and the results derived above would indicate that $d\delta/d\pi$ is positive. However, δ decreases when π is increased from row 1 to row 2. There are two possible explanations for this apparent contradiction. First, although the ratio $(\rho+\pi)/\delta$ begins below the critical value $\sqrt{1-[\tau/(1-k)]}$ in row 1, it ends up above this critical value in row 2. Hence, in raising the rate of inflation from zero to 5 percent, $d\delta/d\pi$ changes sign, and it may be that the negative values of $d\delta/d\pi$ dominate the positive values over this range. Second, Auerbach's table 1 indicates a drop in the real discount rate ρ when going from $\pi=0$ to $\pi=5$ percent. The results derived in this paper have assumed that ρ is constant. However, as shown below, a decrease in ρ will tend to cause a decrease in δ, consistent with the results in Auerbach's table 1.

To calculate the effect of ρ on the depreciation rate δ, simply replace λ_π by λ_ρ and λ'_π by λ'_ρ in (11). Differentiating λ with respect to ρ, we obtain

$$\lambda_\rho = \frac{-(1-\tau)A(\delta)\phi}{(\rho+\delta)^2} - \frac{\tau\delta}{(\rho+\pi+\delta)^2} < 0. \qquad (22)$$

Note that λ_ρ is negative since an increase in the real discount rate reduces the present value of future rentals (including depreciation deductions) to capital.

Differentiating λ' with respect to ρ, we obtain

$$\lambda'_\rho = \frac{-(1-\tau)\phi}{(\rho+\delta)^2}\left[A'(\delta) - \frac{A(\delta)}{\rho+\delta}\right] + \frac{(1-\tau)\phi A(\delta)}{(\rho+\delta)^3} + \tau\frac{\delta-\rho-\pi}{(\rho+\pi+\delta)^3}. \qquad (23)$$

Evaluating this derivative where $\lambda'=0$, we substitute $\tau(\rho+\pi)/(\rho+\pi+\delta)^2$ for $[-(1-\tau)\phi/(\rho+\delta)]\{A'(\delta)-[A(\delta)/(\rho+\delta)]\}$ to obtain

$$\lambda'_\rho = \frac{(1-\tau)A(\delta)\phi}{(\rho+\delta)^3} + \frac{\tau}{(\rho+\pi+\delta)^3}\left(\delta + \frac{\rho+\pi}{\rho+\delta}\pi\right), \qquad (24)$$

which is positive for $\pi \geq 0$. To calculate B_ρ, substitute λ_ρ and λ'_ρ into $B_\rho = -\lambda'_\phi\lambda_\rho + \lambda_\phi\lambda'_\rho$ and use the substitution mentioned in footnote 15 to obtain

$$B_\rho = \frac{(1-\tau)A(\delta)}{(\rho+\delta)(\rho+\pi+\delta)^2}\left\{(q-\tau z)\left[\left(1+\frac{\pi}{\rho+\delta}\right)^2 - 1\right] + \frac{q(q-\tau)}{q-\tau z}\right\}. \qquad (25)$$

Under the reasonable assumption that $\tau < 1-k$, it is the case that $q-\tau>0$ and, hence, $B_\rho>0$. Therefore, since $\lambda'_\rho>0$ and $B_\rho>0$, we

determine from the appropriately modified version of (11) that $d\delta/d\rho > 0$. Therefore, as mentioned above, a decrease in ρ tends to decrease the optimal value of δ.

We have shown that the direction of the effect of inflation on durability depends on relative magnitudes of the nominal discount rate and the depreciation rate. We now examine the effect of inflation on the steady-state flow of capital services by examining the effect on ϕ, the marginal product of capital services. Using Cramer's rule, we obtain

$$\frac{d\phi}{d\pi} = \frac{1}{\Delta}\left\{(1-\tau)c''g'\frac{I/p}{\delta}\left[A'(\delta) - \frac{A(\delta)}{\delta}\right]\lambda'_\pi - g'\frac{A(\delta)}{\delta}\lambda'_\delta\lambda_\pi\right\}. \quad (26)$$

Note that the coefficient of λ'_π is positive and that the coefficient of λ_π is negative. Furthermore, recall that λ_π is negative. Therefore, $d\phi/d\pi$ is positive if either $c'' = 0$ or λ'_π is positive. Since ϕ is a decreasing function of the flow of capital services, it is clear that, in the absence of adjustment costs, an increase in the rate of inflation reduces the flow of capital services, even though the durability of capital may increase or decrease.[16]

IV. Concluding Remarks

By replacing Auerbach's proportional consumption function with consumption behavior determined by intertemporally optimizing consumers, we can derive strong results about the direction of the effect of inflation on durability without appealing to a simulation analysis. If, in the absence of installation costs, $(\rho + \pi)/\delta$ (the ratio of the nominal discount rate to the rate of depreciation) exceeds a certain critical value, then $d\delta/d\pi < 0$, so that an increase in the rate of inflation leads to more durable capital; if this ratio is less than the critical value, then $d\delta/d\pi > 0$. As explained in Section III, these results are broadly consistent with Auerbach's simulation results in which $d\delta/d\pi < 0$ since Auerbach's simulations are confined to the region in which $(\rho + \pi)/\delta$ exceeds the critical value. However, for less durable capital, the ratio $(\rho + \pi)/\delta$ may fall short of its critical value, and the model presented here indicates that $d\delta/d\pi$ would be positive. Regardless of whether an increase in the rate of inflation increases or decreases durability, we find that, in the absence of installation costs, an increase in the rate of inflation reduces the steady-state flow of capital services.

[16] For the case in which $c'' = \infty$, sign $d\phi/d\pi$ = sign λ'_π = sign $d\delta/d\pi$. In this case, the steady-state rate of investment is invariant to the rate of inflation. If an increase in π increases (decreases) δ, then the steady-state flow of capital services $[A(\delta)/\delta](I/p)$ is reduced (increased), thereby increasing (decreasing) ϕ.

References

Abel, Andrew B., and Blanchard, Olivier J. "An Intertemporal Equilibrium Model of Saving and Investment." Report 8006, Univ. Chicago, Center Math. Studies Bus. and Econ., February 1980. Forthcoming in *Econometrica.*

Auerbach, Alan J. "Inflation and the Choice of Asset Life." *J.P.E.* 87, no. 3 (June 1979): 621–38.

Eisner, Robert, and Strotz, Robert H. "Determinants of Business Investment." In *Impacts of Monetary Policy,* compiled by the Commission on Money and Credit. Englewood Cliffs, N.J.: Prentice-Hall, 1963.

Gould, John P. "Adjustment Costs in the Theory of Investment of the Firm." *Rev. Econ. Studies* 35 (January 1968): 47–55.

Lucas, Robert E., Jr. "Optimal Investment Policy and the Flexible Accelerator." *International Econ. Rev.* 8 (February 1967): 78–85.

Treadway, Arthur B. "On Rational Entrepreneurial Behavior and the Demand for Investment." *Rev. Econ. Studies* 36 (April 1969): 227–39.

Uncertain Lifetime, Consumption, and Dissaving in Retirement

James B. Davies

University of Western Ontario

This paper asks whether the continued accumulation, or mild dissaving, observed among the retired can be explained by uncertain lifetime. In the absence of annuities, after an initial period influenced by borrowing constraints, under constant relative risk aversion, uncertain lifetime depresses consumption by a proportion increasing with age if the elasticity of intertemporal substitution in consumption is "small." Illustrative computations, based on actual income and survival data, show that plausible elasticities are sufficiently small to give this effect. The reduction in consumption is large enough to explain much of the lack of decumulation by the elderly.

I. Introduction

It has frequently been observed that the elderly either continue to save in retirement or decumulate much more slowly than would be predicted by the crude life-cycle model without bequest motive under certainty. Most recently, Mirer (1979) has found that, controlling for education, in cross section the mean net worth of American couples continues to rise in retirement.[1] When secular growth is taken into

I would like to thank Glenn MacDonald, Nigel Tomes, Geoffrey Carliner, Chris Robinson, and John Whalley, as well as the participants in the Labor Workshop at the University of Western Ontario and a seminar at McMaster University, for helpful comments. The responsibility for any errors or omissions is my own.

[1] Shorrocks (1975) points out that the slope of the age-wealth profile is biased upward by the lower mortality of the wealthy. Mirer (1979) argues that controlling for education largely corrects for this. While it is not clear that explicit corrections based on estimates of differential mortality by social class provide more reliable results (Shorrocks 1975, p. 163), it should be noted that Shorrocks obtained fairly rapid decumulation in retirement for the United Kingdom by using this alternative.

[*Journal of Political Economy*, 1981, vol. 89, no. 3]

account, this implies an increase, a fortiori, for a cohort. Almost as startling are the positive, or only slightly negative, rates of accumulation found in surveys of saving.[2] However, few, even among the old, *say* they are saving for bequest.[3] Recent work suggests that this may be because human as well as nonhuman intergenerational transfers are possible, and many families are at a corner where only the former are desired (see Drazen 1978; Becker and Tomes 1979; Adams, forthcoming). Thus a bequest-motive explanation of the slow dissaving of the elderly is not as attractive as it might seem.

This paper shows that the observed slow decumulation of the elderly may to a large extent be explained by uncertain lifetime. It does so by developing a model in which the effect of this kind of uncertainty on consumption by persons without a bequest motive can be assessed at different ages. Constant relative risk aversion is assumed for simplicity, but an appendix (obtainable from the author) shows that the results also hold under constant absolute risk aversion. Nonpension annuities are ignored on the argument that a no-bequest life-cycle model is consistent with their observed unimportance in household portfolios only if available annuities are not sufficiently attractive to dominate regular saving instruments for most consumers. With the most plausible taste parameters and rates of return, in this model in the absence of pensions uncertain lifetime will not only depress consumption at all ages but will also have an increasingly severe proportional impact beyond middle age. Illustrative computations, based on actual earnings and survival data, show that this outcome is also likely with realistic pensions and that the effects are sizable. On conservative assumptions, uncertain lifetime may more than halve the mean rate of decumulation among the retired.

The paper is organized as follows. Section II reviews previous work

[2] In three major studies, the following ratios of mean saving to mean net worth for those over 65 were found:

Source	Country	Year	Rate (%)
Lydall (1955, pp. 143, 147)	U.K.	1952	−1.3
Projector and Weiss (1966, p. 110) and Projector (1968, p. 215)	U.S.	1963	+.2
Statistics Canada (1973*a*, p. 84; 1973*b*, p. 140)	Canada	1970	+.8

[3] Only 4 percent of the respondents to the 1962 *Survey of Financial Characteristics* in the United States cited "providing an estate" as a saving objective (Projector and Weiss 1966, table A30). The 1964 Brookings Survey of affluent families (incomes above $10,000) found only 23 percent (of all ages) who were saving to make a bequest (Barlow, Brazer, and Morgan 1966, p. 198).

on the impact of uncertain lifetime on consumption. Section III presents the model, Section IV the illustrative computations.

II. Previous Work

Yaari (1965) studied the impact of uncertain lifetime on the planned rate of change of consumption, examining four cases defined by the presence or absence of a bequest motive and actuarially fair insurance. Except in the cases with a bequest motive, insurance takes the form of either life-insured net borrowing or annuities. Since under actuarial fairness the interest rate on the latter exceeds that on "regular" savings, consumers without a bequest motive hold their entire wealth in the form of annuities.[4]

Omitting a bequest motive, Yaari assumed that a consumer aged t, with probability $P(\tau \mid t)$ of surviving to age τ and a maximum lifetime of T, would have the expected utility integral:

$$V = \int_t^T P(\tau \mid t)\alpha(\tau)U[C(\tau)]d\tau \tag{1}$$

where $C(\tau)$ is consumption at τ, $\alpha(\tau)$ is a subjective discount factor, and the rate of time preference is $1 - P(\tau \mid t)\alpha(t)$.

Maximizing (1) subject to the appropriate constraints yields the following results. In the absence of insurance, since the rate of time preference is higher than under certainty $[P(\tau \mid t) < 1]$, consumption tends to grow more slowly throughout the life cycle. However, actuarially fair insurance makes the price of future consumption $P(\tau \mid t) \exp[-r(\tau - t)]$, where r is the (constant) rate of interest, that is just low enough to offset completely the higher time preference. The result is that, with insurance, consumption grows at about the same rate under uncertainty as under certainty.[5]

It is important to note that Yaari's conclusions concerned only the *shape* of the lifetime consumption path and not its level. Levhari and Mirman (1977, referred to as LM below) have extended Yaari's work by analyzing levels in the absence of insurance. They assume constant rates of subjective discount and relative risk aversion, giving:

[4] The rate of return on an annuity of a year's duration would equal the rate on regular savings plus the probability of death during that year (see Yaari 1965, p. 144).

[5] The actual results are that $\dot{C}(t) = -[r + \hat{\alpha}(t)](U'/U'')$ under both certainty and uncertainty with insurance, while $\dot{C}(t) = -[r + \hat{\alpha}(t) - \hat{P}(t \mid t)](U'/U'')$ under uncertainty without insurance. (A dot denotes a time derivative and a hat a proportional rate of change with time; U' and U'' are first and second derivatives of U with respect to $C[t]$.) Since $C(t)$ will differ in the three cases, and U'/U'' is in general not constant, we can say only, loosely, that under uncertainty consumption *tends* to grow at the same rate as under certainty with insurance but more slowly without.

$$V = \int_t^T P(\tau \mid t) \exp[-\rho(\tau - t)] \frac{C(\tau)^{1-\gamma}}{1-\gamma} d\tau. \qquad (2)$$

Here $-\gamma$ is the elasticity of marginal utility with respect to consumption, and γ is the index of relative risk aversion. The elasticity of intertemporal substitution in consumption, σ, is equal to $1/\gamma$.

Maximizing (2) when resources are composed exclusively of nonhuman wealth, LM find that the impact of uncertain lifetime is a priori indeterminate. In the special case where $r = \rho = 0$, initial consumption will be higher than under certainty if $\gamma < 1$ and lower if $\gamma > 1$. When r and ρ are nonzero, there are a number of distinct cases. It is argued in Section IV of this paper that the most "realistic" is where $\gamma > 1$. In this case, unless r or ρ (or both) is very high, increased uncertainty always leads to a decline in initial consumption.[6] As shown below, this conclusion does not extend to the case where resources are largely in "human" form.

Barro and Friedman (1977) have drawn attention to the fact that, under (2), with actuarially fair insurance, consumption grows at the same rate under uncertain lifetime as under certainty. They have also pointed out that if earnings did not change with age, and r and ρ were zero, consumption would be constant throughout life *at the same level* under certainty and uncertainty. This special result appears to be the only available analysis of the impact of uncertain lifetime on the level of consumption, in the presence of insurance.[7]

These previous contributions might be criticized for focusing either on a world without insurance and annuities or on one in which they are actuarially fair. Both situations are clearly "unrealistic." Must we, then, model insurance with an equilibrium load to make further progress? In the present context, the answer is no. Aside from considerations of analytical difficulty, the exercise is not necessary if we wish to discover whether the life-cycle model without a bequest motive can explain the slow dissaving of the old. Nonpension annuities are in fact of little importance for savers of any age.[8] There are two possible

[6] Levhari and Mirman 1977, p. 275. Note that LM did not regard this case as especially significant. In fact, in their abstract they state "the major result is that if the utility function is Cobb-Douglas and the rate of return is not too large relative to the amount of future discounting then lifetime uncertainty will always increase consumption," which refers to the case where $\gamma = 1$.

[7] An appendix to this paper which analyzes the case with actuarially fair insurance is available from the author. The illustrative computations show that, although, theoretically, a positive or negative impact of uncertain lifetime on consumption is possible, a slight negative impact is likely at all ages. There is no tendency, however, for this to be more severe for the elderly.

[8] In the 1962 *Survey of Financial Characteristics* only 1 percent of households reported any private annuities, and on average these composed less than 0.1 percent of the household portfolio. Surprisingly, the figures are not any higher for those aged 65+ (Projector and Weiss 1966, tables A8 and A31).

explanations. Either there is a widespread strong desire to make bequests, or insurance markets fail to provide annuities sufficiently attractive to outweigh the greater transactions costs and inconvenience of saving in this form.[9] Thus the only case in which the life-cycle model without a bequest motive *could* explain the low consumption of the old is that where regular saving instruments generally dominate annuities. Accordingly, for analytical simplicity it is assumed here that regular savings always dominate annuities.

III. The Impact of Uncertain Lifetime at Different Ages

The object of this paper is to determine the impact of uncertain lifetime on consumption at different ages. One might suppose the appropriate procedure would be to compare complete lifetime plans under certainty and uncertainty. Such a comparison would not, however, tell us the impact of uncertainty on an individual of arbitrary age with given current wealth and other characteristics. It would indicate only differences in projected consumption at various ages for individuals with given wealth at the *beginning* of their lifetimes.[10]

The impact of uncertain lifetime on consumption by an individual of age t with wealth, $W(t)$, and a noninvestment income stream, $E(\tau)$, $t \leqslant \tau \leqslant T$, can be measured by the difference between what he would consume if his lifetime were certain, $C^c(t)$, and if it were uncertain, $C^u(t)$. (The argument t will be omitted in the text below when no ambiguity arises.) In making this comparison we must vary only the degree of uncertainty. In particular, the expected lifetime,

$$\mu(t) = \int_t^T P(\tau \mid t) d\tau + t, \tag{3}$$

must be held constant.

Consumption under Certainty

Under certainty a consumer of arbitrary age, t, would maximize (2) subject to the constraint that discounted consumption up to $\mu(t)$, $\bar{K}^c(t)$, could not exceed total current resources, $R^c(t)$. Resources would be composed of current wealth, $W(t)$, and the discounted value of nonin-

[9] There is a significant problem of adverse selection since survival chances clearly differ widely but may be quite difficult for an insurance company to ascertain.

[10] As mentioned previously, Yaari (1965) showed that under uncertainty consumption tends to grow less quickly than under certainty. Thus, irrespective of the difference in initial consumption (say, at age 20), by retirement, consumption under a complete lifetime plan will almost certainly be lower under uncertainty. This has no bearing, however, on whether an individual aged, say, 65, of given wealth, consumes more or less than he would under certainty.

vestment income, $E(\tau)$, $t \leqslant \tau \leqslant \mu(t)$, denoted $H^c(t)$. Adding the terminal condition $W[\mu(t)] = 0$, we have the constraint:

$$W(t) + H^c(t) - \int_t^{\mu(t)} C^c(\tau) \exp\left[-r(\tau - t)\right]d\tau = 0, \tag{4}$$

or

$$R^c(t) - \bar{K}^c(t) = 0.$$

This is a standard problem in optimal control. Consumption will grow according to

$$\hat{C}^c(\tau) = \frac{r - \rho}{\gamma} = g, \tag{5}$$

where the hat denotes a proportional rate of change with time (as it will throughout). Equation (5) has the obvious solution:

$$C^c(\tau) = \exp\left[g(\tau - t)\right]C^c(t). \tag{6}$$

Substituting (6) into (4) we obtain:[11]

$$C^c(t) = \left\{\int_t^{\mu(t)} \exp\left[(g - r)(\tau - t)\right]d\tau\right\}^{-1} R^c(t) = PC^c(t)R^c(t). \tag{7}$$

The term PC^c, the average "propensity to consume" out of resources, is just $PC^c(t) = [\bar{K}^c(t)/C^c(t)]^{-1} = K^c(t)^{-1}$, that is, the inverse of discounted lifetime consumption normalized by initial consumption. It is helpful to view PC^c in this light since it is usually easy to see how a change in uncertainty affects "normalized" lifetime consumption. The latter simply reflects the shape of the consumption path and the strength of discounting.

Consumption under Uncertainty

Under uncertain lifetime, the consumer maximizes (2) subject to the constraint that wealth may never become negative. (To lend without security to a person whose lifetime is uncertain would be to provide a form of insurance.) As Yaari pointed out, under this constraint the optimal plan consists of a series of intervals where the constraint is effective and others where it is not. I refer to these as "blocked" and "free," respectively. In the LM analysis, where all resources are in the form of an initial stock of nonhuman assets, the entire lifetime is a free interval. Here matters are more complicated.

Over blocked intervals we have, of course, $C^u = E$. In addition, it is clear that the consumption plan is continuous at the beginning, t^*,

[11] Eq. (7) is not simplified, in order to emphasize its similarity to the corresponding expressions under uncertainty, eqq. (10) and (13).

and end, $t^* + m$, of any free interval, with $C^u(t^*) = E(t^*)$ and $C^u(t^* + m) = E(t^* + m)$.[12] Over the course of a free interval, consumption changes according to

$$\hat{C}^u(\tau) = \frac{1}{\gamma}[r - \rho + \dot{P}(\tau)] \tag{8}$$

(Yaari 1965, p. 143).[13] That is, consumption grows less quickly than under certainty because expected utility from future consumption declines without any corresponding drop in price. Solving (8), we obtain:

$$C^u(\tau) = P(\tau \mid t)^{1/\gamma} \exp[g(\tau - t)]C^u(t), t^* \leq t \leq \tau \leq t^* + m. \tag{9}$$

Recalling that wealth must be exhausted over the interval, we have:

$$C^u(t^*) = \frac{\int_{t^*}^{t^*+m} E(t) \exp[-r(t - t^*)]dt}{\int_{t^*}^{t^*+m} P(t \mid t^*)^{1/\gamma} \exp[(g - r)(t - t^*)]dt}. \tag{10}$$

Equations (9) and (10) completely specify the free interval when $t^* + m = T$. In the case where $t^* + m < T$, we have a further relationship. Since $C^u(t^*) = E(t^*)$ and $C^u(t^* + m) = E(t^* + m)$, given (9), $E(t^*)$ and $E(t^* + m)$ are related as follows:

$$E(t^*) = \frac{E(t^* + m)}{P[(t^* + m) \mid t^*]^{1/\gamma} \exp(gm)}, m < T - t^*. \tag{11}$$

An indication of the range of plans that may be obtained with realistic E and P paths and plausible values of r, ρ, and γ is provided by figure 1. The plans shown are the result of the illustrative computations of the next section. Note that in all cases life begins with a blocked interval and there is just one free interval.[14] Part A indicates that with $E = 0$ beyond age 65 (no pensions), the free interval would extend until T. That is, wealth would not be run to zero as long as there was a nonzero probability of surviving any longer. In part B, the consequences of realistic pensions are shown. With a high γ, the consumption path is relatively flat (see [9]) and a free interval lasting

[12] If $C^u(t^*) < E(t^*)$, expected utility could be increased by raising consumption at t^* and reducing it just before. A similar argument holds for $t^* + m$.

[13] Note that, since $P(\tau \mid \tau) = 1$, $\dot{P}(\tau) = \hat{P}(\tau \mid \tau)$.

[14] In general, the consumption plan will not always have these features. The initial blocked interval is caused here by E being low at the outset. A second free interval does not succeed the first because, beyond $t^* + m$, the desired growth rate of consumption is declining because of the rapid fall of P, whereas pensions (if present) fall at a constant rate of 2.25 percent per year. (See the next section for an explanation of this rate.) An increase in the ratio of the growth rates of desired consumption and E would be required for a second free interval to emerge.

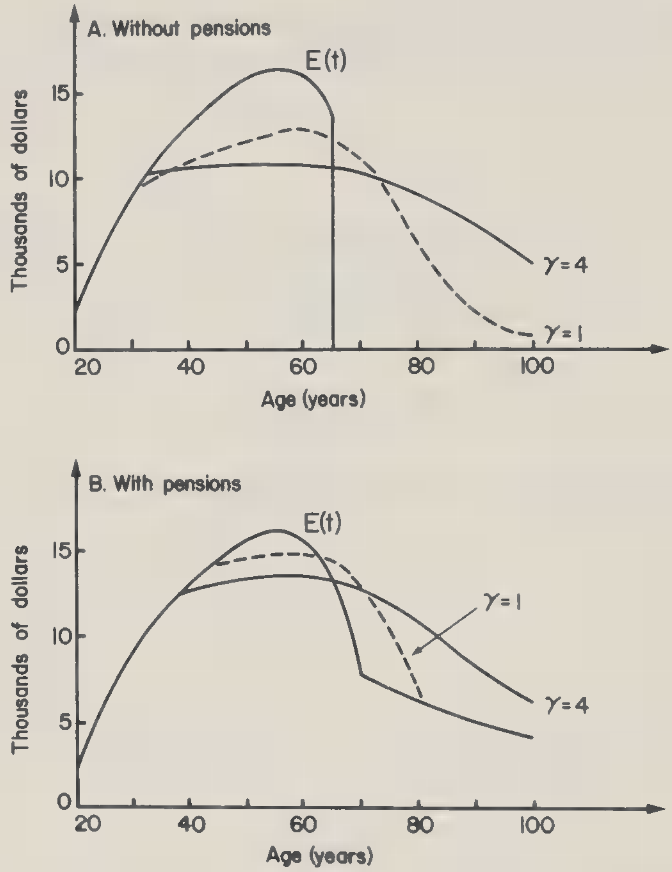


FIG. 1.—Selected optimal consumption paths, $r = .03$, $\rho = .015$

until T is obtained, whereas with $\gamma = 1$ the desired growth rate of consumption declines quickly and a second blocked interval occurs.

Behavior of C^u/C^c *if Free Interval Lasts until* T

How does C^u compare with C^c at different ages? Consider, first, situations where the free interval lasts until T. At any point in the free interval, discounted consumption up to T cannot exceed total "resources," R^u, that is, the sum of current wealth and the present value of the E stream up to T, H^u. Adding the terminal condition $W(T) = 0$, we have:

$$W(t) + H^u(t) - \int_t^T C^u(\tau) \exp\left[-r(\tau - t)\right]dt = 0, \tag{12}$$

or $R^u(t) - \bar{K}^u(t) = 0$. Substituting (9) into (12), we obtain:

$$C^u(t) = \left\{\int_t^T P(\tau \mid t)^{1/\gamma} \exp\left[(g - r)(\tau - t)\right]d\tau\right\}^{-1} R^u(t) = PC^u(t)R^u(t). \tag{13}$$

Again, $PC^u(t) = K^u(t)^{-1}$.

TABLE 1

ANALYSIS OF PC^u/PC^c

g and r	γ	PC^u/PC^c
$g > r$	<1	>1 if γ sufficiently below 1, otherwise <1
	$=1$	<1
	>1	<1
$g = r$	<1	>1
	$=1$	$=1$
	>1	<1
$g < r$	<1	>1
	$=1$	>1
	>1	<1 if γ sufficiently large, otherwise >1

The behavior of C^u/C^c is determined by that of R^u/R^c and PC^u/PC^c. When the free interval lasts until T, in the presence of pensions we have $R^u/R^c > 1$. However, as will be shown, we may have $PC^u/PC^c < 1$, so that $C^u/C^c < 1$ is a possible outcome.

From equations (7) and (13), we see that PC^u/PC^c depends on the relative size of g and r and the value of γ. In Section IV it is argued that γ is typically considerably greater than unity. It is observed that apparently non-credit-constrained households consume at an increasing rate. As long as utility is additively separable over time, this is only consistent with an interest rate greater than the subjective discount rate.[15] Thus realistic restrictions include $\gamma > 1$ and $\rho < r$. Further, if γ is typically considerably above unity, unless ρ is far below zero, which seems unrealistic, we have $g < r$.

Levhari and Mirman have provided a full analysis of the relative values of PC^u and PC^c. Their conclusions are set out in table 1. When $g = r$ and $\gamma = 1$, recalling (3), we find that (7) and (13) imply that $PC^u = PC^c$. Lowering γ below 1 raises $1/\gamma$ and increases PC^u/PC^c, while raising it does the opposite. With $g > r$ and $\gamma = 1$, $PC^u/PC^c < 1$ since greater relative weight is placed on the more remote future in the denominator of PC^u than in that of PC^c. Raising γ simply reinforces this. On the other hand, although reducing γ below 1 tends to increase PC^u/PC^c, the ratio will not rise above unity until γ is sufficiently small, so that a simple result is not obtained.

Now, in the apparently realistic situation where $g < r$, we have $PC^u/PC^c > 1$ with $\gamma \leqslant 1$, again because more weight is placed on the more remote future in the denominator of PC^u. However, raising γ

[15] In Yaari's framework, which does not require constant risk aversion or constant subjective discounting, consumption will not rise unless the rate of return exceeds the subjective rate of discount; see Yaari (1965), pp. 143, 147.

above unity reduces $1/\gamma$ and tends to make PC^u fall. Clearly, with sufficiently large γ we obtain $PC^u/PC^c < 1$. In the next section we find that with actual earnings and survival data, γ does not have to be far above unity for this result to occur.

In order to study changes in PC^u/PC^c with age, one may compare the proportional rates of change:[16]

$$\widehat{PC^c}(t) = PC^c(t)(1 - \exp\{(g - r)[\mu(t) - t]\}\dot{\mu}(t)) + (g - r), \quad (14)$$

and

$$\widehat{PC^u}(t) = PC^u(t) + \frac{\dot{P}(t)}{\gamma} + (g - r). \quad (15)$$

To interpret these relationships, it is helpful to note that $\widehat{PC^i} = -\hat{K}^i$. Consider first a situation where the lifetime was actually fixed and $g = r$. In this case, normalized lifetime consumption, K^i, would simply equal the remaining life expectancy, $\mu(t) - t$. Each year K^i would fall by a percent equal to its own inverse. Hence $PC^i = (K^i)^{-1}$ would *rise* by this same percent, that is, PC^i itself.

When allowing $g \neq r$, say $g < r$, we give K^i some tendency to rise as t increases since weaker discounting (compounding, if $g > r$) is applied. This gives rise to the $(g - r)$ in both (14) and (15). Finally, in fact lifetime is not fixed. Under certainty, the extension in possible lifetime increases normalized lifetime consumption, reducing PC^c, while under uncertainty, the increased probability of survival to any future date also raises K^i and reduces PC^i.

Equations (14) and (15) show that, in the main, whether PC^u/PC^c tends to rise or fall depends on whether it starts out greater or less than unity. When $PC^u/PC^c < 1$, a decline *may* not necessarily occur. However, since $\dot{\mu}(t)$ and $\dot{P}(t)$ are both typically quite small, in our "most realistic" case, of $g < r$ and γ well above unity, it is likely that PC^u/PC^c will be low enough to ensure its continuous decline. Whether this translates into a decline in C^u/C^c depends entirely on the behavior of R^u/R^c.

In the absence of pension income, since most people retire before the age at which death was expected, even from the viewpoint of the start of adult life, R^u is typically the same as R^c. Hence everything said above about PC^u/PC^c applies to C^u/C^c. That is, with the most plausible parameters $C^u/C^c < 1$ and declines monotonically with age over the free interval. In addition, it is clear that $C^u/C^c < 1$ during the initial blocked interval (C^u will be less than the "free" value given by eq. [13]), although the ratio would likely typically *increase* with age prior to t^* since E rises quickly for the young. (Recall $C^u = E$ in the blocked

[16] In deriving (15) let $\pi(\tau)$ = the unconditional probability of surviving to age τ. Then $P(\tau|t) = \pi(\tau)/\pi(t)$, and $\partial P(\tau|t)/\partial t = -\pi(\tau)\pi(t)^{-2}\dot{\pi}(t)$, or $-P(\tau|t)\dot{P}(t)$.

interval.) These conclusions are particularly interesting since many of the studies indicating slow decumulation by the old were conducted when pensions were less sizable and widespread than they are today.

When pension income is present, we of course have $R^u/R^c > 1$; and even when the free interval continues to last until T, the above conclusions will hold only with a higher value of γ. In the next section we find that the required level of γ is not, however, implausible. Here we should note that, although $R^u/R^c > 1$, it has some tendency to fall with age. The R^i has two components, W and H^i, the former of which we hold constant in the comparison of certainty and uncertainty. Since most often W is rising and H^i is falling, R^u and R^c have a basic tendency to become more equal. In the next section we find that in practice this dominates, and R^u/R^c typically falls with age. It should be noted, however, that the behavior of H^u/H^c is in opposition. We have

$$\hat{H}^u(t) = r - \frac{E(t)}{H^u(t)}, \tag{16}$$

and

$$\hat{H}^c(t) = r + \frac{E[\mu(t)] \exp\{-r[\mu(t) - t]\}\dot{\mu}(t) - E(t)}{H^c(t)}. \tag{17}$$

Since $E[\mu(t)] \exp\{-r[\mu(t) - t]\}\dot{\mu}(t)$ is usually small relative to $E(t)$ and $H^u > H^c$, this means that H^u tends to fall more slowly than H^c.

Behavior of C^u/C^c *if Free Interval Ends before* T

As we saw in figure 1, in the presence of pensions the free interval may terminate before T. In this case, it is likely that $C^u/C^c > 1$ throughout the free, as well as the following blocked, intervals.

In the computations of the next section, in cases where the free interval ends before T, we usually find that $\mu(t^*) < t^* + m$. Hence $\mu(t)$ must rise and become equal to $t^* + m$ at some point in the interval. At this age C^u exceeds C^c. In general, we have:

$$C^u(t) = \frac{W(t) + \int_t^{t^*+m} E(\tau) \exp[-r(\tau - t)]d\tau}{\int_t^{t^*+m} P(\tau \mid t)^{1/\gamma} \exp[(g - r)(\tau - t)]d\tau}, \quad t^* \leq t < t^* + m, \tag{18}$$

and

$$C^c(t) = \frac{W(t) + \int_t^{\mu(t)} E(\tau) \exp[-r(\tau - t)]d\tau}{\int_t^{\mu(t)} \exp[(g - r)(\tau - t)]d\tau}, \quad 0 \leq t \leq T. \tag{19}$$

For $\mu(t) = t^* + m$ relevant resources are the same, but the denominator is smaller under uncertainty, and $C^u > C^c$.

What can we say about C^u/C^c at other points? Before the age where $\mu(t) = t^* + m$, the longer horizon under uncertainty makes the denominator of (18) larger relative to that of (19). However, relevant human resources in the numerator of (18) are greater than those in (19). Similarly, above the point where $\mu(t) = t^* + m$, the denominator of (18) falls relative to that of (19) but the numerator does the same. In the illustrative computations, the net result is often that $C^u > C^c$ throughout the free interval.

Finally, after the free interval we are likely to have $C^u/C^c > 1$ again. To see this, consider a case where $t^* + m$ occurs in retirement and the pension changes at a constant percentage rate. As long as g exceeds this rate, which appears likely, positive saving will be desired under certainty, and $C^c < E = C^u$.

IV. Quantitative Impact

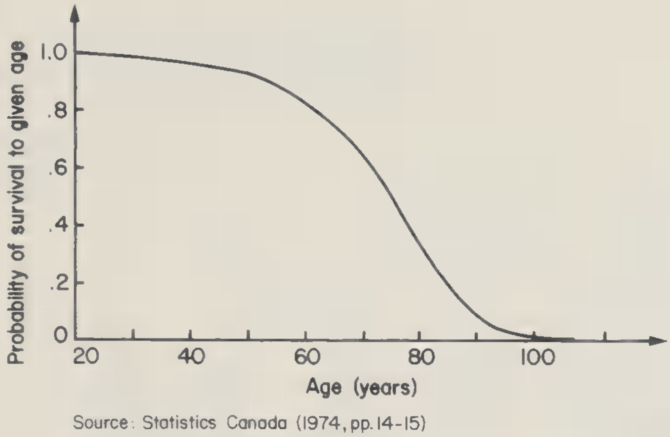
This section provides a quantitative study of the likely impact of uncertainty on consumption at different ages. It does this by computing hypothetical consumption plans under certainty and uncertainty at different ages, using actual data on noninvestment income and survival probabilities and a range of values of r, ρ, and γ. The first step is to compute the complete lifetime consumption plan under uncertainty. This generates a time path for $W(t)$. Next, at intervals of 1 year I ask how much a person would consume, given the $W(t)$ computed in the first step and a *certain* lifetime equal to life expectancy at that age.

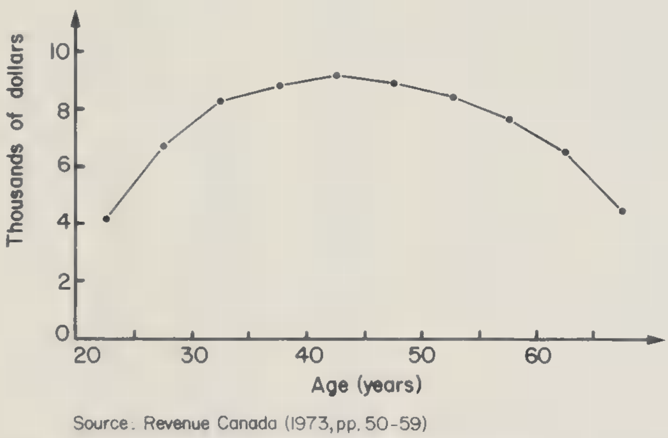
Figures 2 and 3 plot the data used. An inverse logistic curve fits the P path in figure 2 well and allows the use of continuous time in the computations.[17] Similarly, the lifetime path of mean noninvestment income in figure 3 is approximated from ages 20 to 70 by a fourth-order polynomial.[18] Beyond age 70, it is assumed that E declines at the observed trend rate of 4.25 percent per year. It is also assumed that secular growth shifts the entire cross-section profile upward at a rate of 2 percent per year. This explains why the E path for an *individual* shown in figure 1 differs from the cross-section profile of figure 3.

The computations were carried out with alternative values of r, but I focus mainly on the case where $r = .03$, suggested by popular estimates of mean household rates of return (see, e.g., Boskin 1978, p. S19; Feldstein, Green, and Sheshinski 1978, p. S64). The subjective

[17] The curve $P(t) = [(1 + a)\exp(-bt)]/[1 + a\exp(ct)]$ was fitted by least squares, using numerical methods ($R^2 = .999$). Values of the parameters are $a = 1.00031$, $b = -0.002114$, and $c = 0.14651$.

[18] The curve, $E(t) = -36{,}999.4 + 3520.22t - 101.878t^2 + 1.34816t^3 - 0.00706233t^4$, was fitted by ordinary least squares ($R^2 = .999$).





FIGS. 2–3.—Fig. 2, Male survival probabilities (Canada 1971). Fig. 3, Average male noninvestment income (Canada 1971).

rate of discount is limited to the arguably "realistic" range $0 \leqslant \rho \leqslant r$. (It was suggested in the previous section that $\rho > r$ is implausible, while allowing $\rho < 0$ simply reinforces the results obtained below.) The parameter γ is allowed to vary from 0.5 to 5.0, reflecting both a desire to cover the alternatives $\gamma \lesseqgtr 1$ and considerable evidence that γ is typically much above unity.

A variety of evidence on γ is available. Friend and Blume (1975, p. 920) estimated it in the context of risk aversion, using data on household asset holdings, and found that it was likely to be in excess of 2.0. Farber (1978), attempting to infer the preferences of members of the United Mine Workers from the results of collective bargaining in the coal industry, also estimated γ as the index of relative risk aversion, obtaining alternative estimates of 3.0 and 3.7 (p. 935). Further evidence is provided by Ghez and Becker (1975, p. 133) and MaCurdy (1979, p. 34), who, in studies of the allocation of goods and time over the life cycle, estimated upper bounds on σ of .28 and .33, re-

TABLE 2

C^u/C^c WITHOUT PENSIONS

	γ							
	$r = .03$						$r = .015$	$r = .045$
AGE	.5	1	2	3	4	5	4	4
				$\rho = 0$				
20	1.060*	.441*	.306*	.274*	.260*	.252*	.236*	.285*
30	1.264	1.000	.903	.872	.857	.835*	.794	.886*
40	1.305	1.000	.875	.833	.811	.798	.748	.863
50	1.375	1.000	.832	.775	.746	.728	.684	.800
60	1.498	1.000	.771	.695	.656	.633	.600	.709
70	1.678	1.000	.702	.609	.564	.538	.518	.609
80	1.815	1.000	.668	.571	.525	.499	.490	.559
				$\rho = r/2$				
20	.441*	.306*	.260*	.247*	.241*	.237*	.226*	.257*
30	1.216	.986	.857*	.820*	.802*	.792*	.798*	.810*
40	1.261	1.028	.903	.855	.830	.814	.760	.885
50	1.342	1.032	.862	.798	.765	.744	.695	.823
60	1.481	1.038	.802	.718	.674	.648	.610	.733
70	1.675	1.041	.731	.629	.579	.550	.526	.630
80	1.823	1.039	.691	.586	.537	.508	.496	.576
				$\rho = r$				
20	.224*	.224*	.224*	.224*	.224*	.224*	.217*	.234*
30	.752*	.752*	.752*	.752*	.752*	.752*	.769*	.743*
40	1.064*	1.044	.927	.876	.847	.829	.771	.904
50	1.293	1.054	.889	.820	.782	.759	.705	.845
60	1.446	1.067	.830	.739	.692	.662	.619	.756
70	1.657	1.076	.758	.649	.594	.562	.533	.652
80	1.821	1.074	.714	.602	.548	.517	.502	.593

* Nonnegative wealth constraint effective under uncertainty.

spectively, implying $\gamma \geqslant 3.6$. Finally, recent studies of the elasticity of aggregate saving with respect to the real rate of interest suggest high values of γ as well. Boskin (1978, p. S4), for example, estimates the uncompensated elasticity to lie in the range 0.3–0.4 in the United States. Including pensions, with $r = .03$ and $\rho = .015$, my model implies uncompensated aggregate saving elasticities of 0.9, 0.5, and 0.3 with $\gamma = 3$, 4, and 5, respectively.[19]

To sum up: The available evidence suggests that γ may typically be above 3; that a value of 4 may be a "best guess" at the true figure; and that even $\gamma = 5$ could not be rejected out of hand as implausible.

Table 2 presents the illustrative computations for the case where E is set to zero beyond age 65, that is, where there are no pensions. A value of γ below 1 gives $C^u/C^c > 1$ and rising with age in the free

[19] Alternative values of r and ρ yield similar results. Using $\gamma = 4$, with $r = .06$ and $\rho = .03$ the elasticity is 0.6, while with $r = .03$ and $\rho = .0$ we get a figure of 0.4, e.g.

TABLE 3

C^u/C^c WITH PENSIONS

	γ							
	$r = .03$						$r = .015$	$r = .045$
AGE	.5	1	2	3	4	5	4	4
	$\rho = 0$							
20	.997*	.415*	.288*	.258*	.245*	.237*	.217*	.274*
30	1.346	1.089	.909*	.827*	.790*	.769*	.740*	.836*
40	1.397	1.109	.992	.946	.922	.908	.915	.935
50	1.491	1.145	.988	.922	.888	.867	.886	.899
60	1.669	1.211	.983	.885	.835	.805	.842	.840
70	1.965	1.337	.994	.844	.772	.730	.795	.766
80	1.808	1.500	1.065	.843†	.748†	.696†	.793†	.726†
	$\rho = r/2$							
20	.415*	.288*	.245*	.232*	.227*	.223*	.207*	.247*
30	1.232*	.909*	.790*	.756*	.739*	.730*	.714*	.764*
40	1.330	1.168	1.043*	.974	.945	.926	.929	.960
50	1.437	1.178	1.030	.959	.917	.891	.904	.933
60	1.618	1.266	1.046	.937	.874	.835	.865	.888
70	1.705	1.397	1.090	.917	.822	.767	.825	.827
80	1.222*	1.230	1.193	.943†	.807†	.737†	.830†	.798†
	$\rho = r$							
20	.210*	.210*	.210*	.210*	.210*	.210*	.199*	.224*
30	.693*	.693*	.693*	.693*	.693*	.693*	.688*	.701*
40	.935*	.935*	.935*	.935*	.935*	.935*	.943	.925*
50	1.138*	1.138*	1.056	.991	.944	.914	.921	.964
60	1.398*	1.263	1.089	.985	.912	.866	.889	.934
70	1.123*	1.318	1.154	.992	.875	.807	.857	.895
80	1.088*	1.088*	1.220	1.053	.876†	.784†	.871†	.887†

* Nonnegative wealth constraint effective under uncertainty.
† No blocked interval succeeds free interval.

interval, while $\gamma > 1$ yields the opposite, as predicted in the previous section. With $\gamma \geqslant 3$ dramatic results are obtained. Using the "best guess" figure of $\gamma = 4$ and the combination $r = .03$ and $\rho = .015$, for example, maximum negative impacts on consumption are 23.5 and 46.3 percent for the middle-aged and old, respectively. With the same, quite conservative, parameters, the mean ratio of dissaving to net worth for those aged 65–85 declines from 9.0 to 3.7 percent with the introduction of uncertainty. Note, finally, that the results are robust with respect to changes in r.

Table 3 shows that when noninvestment income in retirement is introduced, the ratio C^u/C^c rises in most cases. However, with $\gamma \geqslant 3$ and $\rho < r$, the impact of uncertainty remains negative and much

larger for the retired than for the middle-aged. When we focus again on the $r = .03$, $\rho = .015$, $\gamma = 4$ case, we see that, while uncertainty reduces consumption in middle age by up to 8.3 percent, it depresses that in retirement by up to 19.3 percent. Although this is a more modest effect than obtained without pensions, it is still impressive, as shown by the fact that with the same parameter values the average rate of decumulation between ages 65 and 85 falls from 7.0 to 2.9 percent with the introduction of uncertainty.[20] Furthermore, isolation of the $r = .03$, $\rho = .015$, $\gamma = 4$ case is conservative. Small changes in parameter values can make the effect even more sizable. By reducing ρ to 0 or by raising γ to 5, for example, we increase the maximum negative impact on consumption for the elderly to about 25 percent and reduce the mean rates of decumulation under uncertainty for those aged 65–85 to 2.4 and 2.1 percent, respectively. Finally, note that the results are again robust with respect to r.[21]

V. Conclusion

This paper has shown that the life-cycle model without a bequest motive can help to explain one of the most interesting features of actual consumption behavior, the typical low rate of dissaving among the elderly, when uncertain lifetime is taken into account. Theory shows that, with the most plausible parameters, a negative impact of uncertain lifetime which increases in proportional severity with age is likely in the absence of pensions and may also occur when pensions are present. Illustrative computations based on actual income and survival data show that this effect is, in fact, obtained with or without pensions. Furthermore, the magnitude of the impact is sufficient to suggest that uncertain lifetime could provide the major element in a complete explanation of the slow decumulation of the retired.

References

Adams, James D. "Personal Wealth Transfers." *Q.J.E.* (forthcoming).

Barlow, Robin; Brazer, Harvey E.; and Morgan, James N. *Economic Behavior of the Affluent.* Washington: Brookings, 1966.

[20] The contrast in decumulation rates remains so striking because a given proportional difference in consumption implies a *larger* proportional difference in saving when $E > 0$ because the scale of private wealth holding and dissaving is reduced.

[21] White (1978, p. 559) states that she (1976, chap. 5) demonstrated that the no-bequest life-cycle model without insurance cannot explain aggregate personal saving in the United States. Indeed, the simulations of White (1976, p. 115) show consumption increasing as a result of uncertain lifetime in all cases considered. The explanation of the apparent conflict with the present paper is that White examined only values of γ between 0.5 and 1.5. Thus her results correspond to those shown here in the initial columns of table 3.

Barro, Robert J., and Friedman, James W. "On Uncertain Lifetimes." *J.P.E.* 85, no. 4 (August 1977): 843–49.

Becker, Gary S., and Tomes, Nigel. "An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility." *J.P.E.* 87, no. 6 (December 1979): 1153–89.

Boskin, Michael J. "Taxation, Saving, and the Rate of Interest." *J.P.E.* 86, no. 2, pt. 2 (April 1978): S3–S27.

Drazen, Allan. "Government Debt, Human Capital, and Bequests in a Life-Cycle Model." *J.P.E.* 86, no. 3 (June 1978): 505–16.

Farber, Henry S. "Individual Preferences and Union Wage Determination: The Case of the United Mine Workers." *J.P.E.* 86, no. 5 (October 1978): 923–42.

Feldstein, Martin S.; Green, Jerry; and Sheshinski, Eytan. "Inflation and Taxes in a Growing Economy with Debt and Equity Finance." *J.P.E.* 86, no. 2, pt. 2 (April 1978): S53–S70.

Friend, Irwin, and Blume, Marshall E. "The Demand for Risky Assets." *A.E.R.* 65 (December 1975): 900–922.

Ghez, Gilbert R., and Becker, Gary S. *The Allocation of Time and Goods over the Life Cycle.* New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1975.

Katz, Eliakim. "A Note on Uncertain Lifetimes." *J.P.E.* 87, no. 1 (February 1979): 193–96.

Levhari, David, and Mirman, Leonard J. "Savings and Consumption with an Uncertain Horizon." *J.P.E.* 85, no. 2 (April 1977): 265–81.

Lydall, Harold F. "The Life Cycle in Income, Saving, and Asset Ownership." *Econometrica* 23 (April 1955): 131–50.

MaCurdy, Thomas E. "An Empirical Model of Labor Supply in a Life Cycle Setting." Nat. Bur. Econ. Res. Working Paper no. 421, December 1979.

Mirer, Thad W. "The Wealth-Age Relation among the Aged." *A.E.R.* 69 (June 1979): 435–43.

Projector, Dorothy S. *Survey of Changes in Family Finances.* Washington: Board Governors, Federal Reserve Board, 1968.

Projector, Dorothy S., and Weiss, Gertrude S. *Survey of Financial Characteristics of Consumers.* Washington: Board Governors, Federal Reserve Board, 1966.

Revenue Canada. *1973 Taxation Statistics.* Ottawa: Minister Supply and Services, 1973.

Shorrocks, A. F. "The Age-Wealth Relationship: A Cross-Section and Cohort Analysis." *Rev. Econ. and Statis.* 57 (May 1975): 155–63.

Statistics Canada. *Family Expenditure in Canada 1969.* Publication no. 62-535. Ottawa: Ministry Indus., Trade, and Commerce, 1973. (*a*)

———. *Incomes, Assets and Indebtedness of Families in Canada.* Publication no. 13-547. Ottawa: Ministry Indus., Trade, and Commerce, 1973. (*b*)

———. *Life Tables: Canada and Provinces.* Publication no. 84-532. Ottawa: Ministry Indus., Trade, and Commerce, 1974.

White, Betsy Buttrill. "On the Rationality of Observed Saving: A Critique of the Life Cycle Hypothesis." Ph.D. dissertation, Stanford Univ., 1976.

———. "Empirical Tests of the Life Cycle Hypothesis." *A.E.R.* 68 (September 1978): 547–60.

Yaari, Menahem E. "Uncertain Lifetime, Life Insurance, and the Theory of the Consumer." *Rev. Econ. Studies* 32 (April 1965): 137–50.

What Kind of a Science Is Economics? A Review Article on *Causality in Economics* by John R. Hicks

Christopher A. Sims

University of Minnesota

As computational costs have declined, economists have used increasingly complicated statistical models in analyzing business cycles. Most of these models have been based on extensions of the IS-LM framework developed by Hicks, combined with the simultaneous equations statistical framework which was the central subject matter of econometrics. These large econometric models were, and still are, successful in certain important respects. By making possible the systematic use of large amounts of statistical information and the codification of forecasting and projection procedures, they provided good forecasts—as good as or better than averages of "judgmental" forecasters. The models and their services have become commercially valuable. Nonetheless, they seem to many economists to have outgrown their underpinnings in economic and statistical theory.

The economic theory entering the models does not in itself explain the existence of recurrent business cycles. In practice, the fitted models usually explain much of what might be labeled "the character of the business cycle" as a consequence of properties of the "error terms" in the model about which theory has little to say. This is not entirely the fault of the economic theory. The simultaneous equations statistical theory makes heavy demands in this area of application on supposed a priori knowledge. It is therefore perhaps not surprising that the actually available economic theory leaves large gaps to be filled with conventional or ad hoc assumptions. The burden of ad hoc assumptions and implicit statistical procedures carried by the models for these reasons weakens them somewhat as tools for forecasting and

[*Journal of Political Economy*, 1981, vol. 89, no. 3]

policy projection. It weakens them still more as scientific instruments when important economic issues are subject to dispute.

Reflecting these sources of dissatisfaction are two current developments in macroeconomic research. On the one hand, some economists are building stochastic theoretical models which have testable probabilistic implications for the behavior of major aggregate time series as stochastic processes. On the other hand, econometricians have been looking at methods for reducing the ad hockery of the standard methodology by formalizing the procedures for controlling the dimensionality of models and for accounting for drift in model structure.

These developments seem to me natural and hopeful. They do, though, point to an increasingly complicated and intense interaction between theory and data, using probabilistic models. What is going on is not quite like what has gone on in any other science, though one can see some analogies. Is what is going on useful? What are the criteria for confirming or disconfirming models of this type? Are the formal and informal standards for reporting results in this area reasonable?

According to its introduction, the book under review grew out of Hicks's meditation on a 1974 conference on the microfoundations of macroeconomics, which he attended and saw as a failure. It also reflects his view that "econometrics is now in some disarray." The book discusses scientific method in general and then in this light the special character of economics, especially macroeconomics, as science. As is evident from the book's length (124 + xii pages), the argument is concise, and it is presented in a conversational tone. The central conclusions, as I see them, are that "as economics pushes on beyond 'statics' it becomes less like science, and more like history" (p. xi) and that "the probability calculus is no excuse for forgetfulness" (p. 122).

To explain completely what Hicks means by these phrases would require more space than I have in this review, but some explication is obviously required. The former idea is that dynamic economic theories must inherently be incomplete, imprecise, and therefore subject to variation over time. One reason for this is that economic cause-effect relations involve a "recognition delay" about which theory has little to say and which may be expected to be variable. For example, when a price change causes a change in an amount purchased, there will be a delay between the actual price change and consumers' hearing of it, as well as a possible further delay before consumers decide to act on the information. It is wrong, then, to expect economic theories to be complete, mechanical, and divorced from reference to specific historical circumstances.

The latter idea is that, particularly in economic time series, probability models are often inappropriate because the data are interde-

pendent in ways which make treating them as "random" patently unjustifiable. Hicks argues that probability judgments in economics, both in microeconomic decision making and in evaluation of competing economic theories, often result in incomplete orderings, with many possibilities lying in a gray zone of noncomparable probabilities. Thus the usual axiomatic arguments for the probability calculus are unjustified in a wide range of economic applications. He cites as an example of this objection to the probability calculus in inference the situation in which a new fact which seems not to fit existing theory emerges, and a new but ad hoc theory is put forward to "explain" it. Because a priori and likelihood considerations point in different directions in this case, he argues that the two theories lie in the gray zone of noncomparability and that the reasonable action to take is to reserve judgment. Similar considerations, Hicks argues, suggest that economic behavioral models of decision making under uncertainty ought not to be formulated taking the usual probability calculus for granted. Probabilistic models and methods of inference are often awkward and misleading tools in application to economics.

Hicks's discussion is valuable for the insights it presents into the special character of economics as a science. Its conclusions, though, seem to me deeply mistaken, in part because the argument too easily resolves some of the problems it raises.

My dissatisfaction with the Hicks book is easier to explain if we compare it with another recent book on a closely related subject—*Chance, Cause, Reason*, by Arthur W. Burks (1977). The Burks book is expansive (almost 700 pages), where Hicks's is concise; it is densely formal in parts, where Hicks's is persistently informal; it very seldom gives the impression of having chosen an easy resolution of a difficulty, where Hicks's does. Burks, for example, considers at some length the objection to the probability calculus given central place by Hicks: that probability judgments do not yield a complete ordering. Burks argues that these objections are correct in principle but that they are in the nature of qualifications to the range of applicability of the calculus—when no decision has to be made, or when the decision to be made is unimportant, then it may not be "economic" (to use Burks's word) to rank all probabilities and utilities. But (this is my conclusion, not Burks's) in important issues of scientific inference or in economic decision making which involves substantial amounts of money and utility, it is both normatively and descriptively reasonable to treat probabilities and utilities as completely ordered. In taking decisions, we implicitly rank them in any case.

I could go on citing Burks in rebuttal to parts of Hicks's argument with which I disagree, but it would be better for the reader of this review to read Hicks and struggle with Burks (I for one cannot "read"

Burks in the ordinary sense of the term) and reach his own conclusions. For despite my sympathy with Burks's style and with many of his arguments, I do not see his book as in any sense a substitute for or a refutation of Hicks. Burks is constructing a model of scientific inference, and it seems that the hard sciences are the focus of his interest. He rightly points to the need for some degree of uniformity in scientific laws over space and time if inference is to be possible and also rightly (in my view) claims that his model of inference applies even if there is some variation of laws in space and time. He rightly claims that there is a sense in which empirical probability statements (e.g., "the probability that this coin will come up heads is .5") can be true, and from this point builds up an intersubjective (as opposed to Savage's personalistic) theory of inference. Economists do often aim at influencing professional consensus as to what causal structure plausibly underlies historical data. If we are to think about how economists do this or ought to do this, an intersubjective theory like Burks's fills an important gap in Savage's subjective theory.

But Hicks is right about some of the central differences between economics and other sciences—temporal instability of our theories is not a marginal qualification to our methods of inference but a central aspect of our difficulties in determining what is true. And treating "empirical truth" as a primitive notion seems too simple to lead to a theory applicable to economics. The problem of the "ad hoc theory," arising from the dearth of "empirically true" probability statements in economics and the near impossibility of generating them by experiment, is central to us. Burks's system, while relevant to economics, fascinating on its own terms, and stimulating as an example, does not confront in detail some of the central problems of inference in economics.

Hicks is not explicit about how his broad conclusions apply to specific types of current work in economics. (Indeed, except for the author's own works, no publications in macroeconomics, probability, or statistics less than 15 years old are cited.) He does in his final paragraph appear to condemn all use of probability methods on cross sections not generated by random sampling:

> Thus it is not at all sensible to take a small number of observations (sometimes no more than a dozen observations) and to use the rules of probability theory to deduce from them a "significant" general law. For we are assuming, if we do so, that the variations from one to another of the observations are random, so that if we had a larger sample (as we do not) they would by some averaging tend to disappear. But what nonsense this is when the observations are derived, as

> not infrequently happens, from different countries, or localities, or industries—entities about which we may well have relevant information, but which we have deliberately decided, by our procedure, to ignore. [Pp. 121–22]

I think Hicks is wrong here. Any use of experience to reach better decisions must involve "deliberately ignoring" information, whether the experience is the historical record or a sequence of repeated experiments. In physical experiments there is always observational error, which is treated as random. A good experimenter may have many bits of information he judges relevant to explaining the apparent random error in a sequence of experiments. In writing up the experiments, though, he does not report which research assistant was on duty at each meter reading, together with an analysis of each assistant's bias and mean square deviation as a meter reader. In physical sciences there is substantial consensus as to what is to be treated as random.

In social sciences, with history the most extreme instance, there is much less consensus. It is nonetheless necessary to ignore some information, to treat it as unsystematic or random, if the past is to be useful. Because there is inevitably room for dispute about what is systematic and what is not in analyzing history, historians do not dogmatically suppress discussion of information which some might consider "random" or unsystematic. Nonetheless, when a decision has to be made, if history is to be relevant, some choice must be made as to what is random. A political decision maker might well need to judge how likely it is that a Marxist revolution in his country would be followed by political democracy. He might think that the historical record concerning previous Marxist revolutions is relevant to this question. However, it is obvious that every previous revolution has special characteristics which are well known and which are potentially relevant to explaining whether the postrevolutionary government emerged as politically democratic. In applying the historical record to his own problem, the decision maker will have to treat at least some of these special characteristics as random.

The same type of argument justifies, in clinical trials of a new drug, reporting statistical summaries of the outcome of therapy rather than, case by case, the best guess of the attending physician about why the result of therapy was what it was for this case. And the same type of argument again justifies using statistical methods for observation on countries or regions even when we could adduce a great deal of special information about each observation which might possibly help explain its "random" component.

Economists must inevitably try to sort out systematic patterns from

random variations in the past—if only because, unassisted, policy-makers would do the same thing more naively. In doing so economists will need probabilistic models and statistical methods of inference. Like historians, though, they must accept that a single agreed view of the causal structure of the record they examine will never emerge. Perhaps Hicks's stimulating but unsatisfying book will lead someone to work out a philosophy of inference which confronts both these aspects of economics as science.

References

Burks, Arthur W. *Chance, Cause, Reason: An Inquiry into the Nature of Scientific Evidence.* Chicago and London: Univ. Chicago Press, 1977.
Hicks, John R. *Causality in Economics.* New York: Basic, 1979.

Comments

To Save or Savor: The Rate of Return to Storing Wine

Elizabeth Jaeger

University of Virginia and Freemark Abbey Winery

For \$28,000 one can buy a Jaguar XJ6, 308 shares of McDonnell Douglas stock, a one-tenth interest in an Oklahoma gas well, 4½ years of undergraduate life at Stanford, or one bottle of Chateau Lafite, vintage 1806. "That's \$1,166.67 an ounce. At this moment gold is worth only \$275 an ounce, and you can't smell or drink it" (*Saturday Review of Literature* 1979, p. 38). Nine years ago, \$70 could persuade a Paris restaurant to part with this same bottle of liquid gold (Goldwyn 1979, p. 4).

But for balance admire the growth of a slightly more pedestrian vintage, a 1959 Beaulieu estate-bottled Cabernet. Today a case of it may be purchased for \$1,200. For the more modest budget, \$950 will buy a case of 1970 Beaulieu George Latour Private Reserve Cabernet. All this goes to show that wine, the right wine, in the last 10 years has provided the prudently daring investor with a spectacular return.

The virtues of wine as a liquid asset have been tested recently by Krasker (1979). Krasker cannot reject a hypothesis that wine offers no extraordinary rate of return; the wine investor can expect no greater return than that which accrues to riskless assets. While Krasker finds the concept that an investment in wine outperforms Treasury bills in nominal return "incompatible with the way economists believe most assets are priced," his result seems counter to intuition. Wine as a subcategory of riskless assets seems contrary to wine's very nature.

Wine is a living substance. Its maturation process extends well beyond the bottling day, lasting as long as 20–40 years in the case of some reds, while a few others refuse to retire at 65. But *the* proper age

[*Journal of Political Economy,* 1981, vol. 89, no. 3]

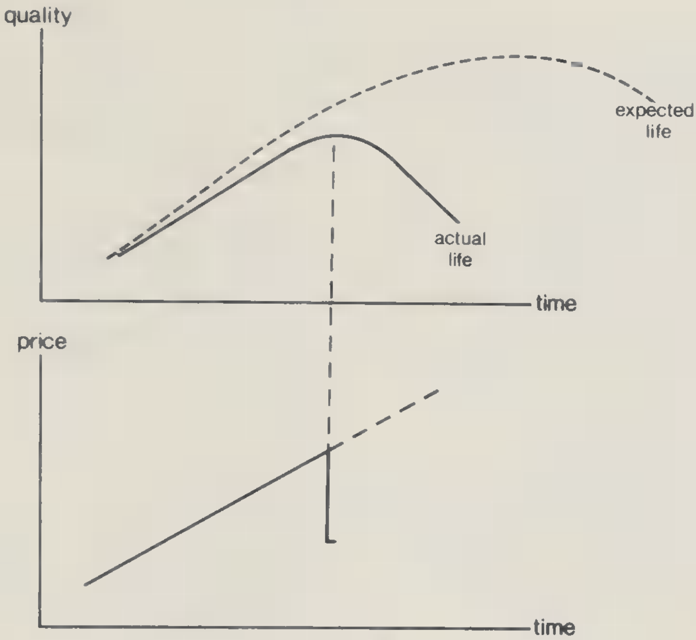


FIG. 1

at which to consume a wine is not some known date but rather a peak which must be discovered by the bulk purchaser through experimentation. When frequently posed with the question when to drink, Harry Waugh, the famous wine writer, would respond, "Just imagine looking around a room full of people and somebody saying to you, 'Tell me at what date each of these individuals is going to die' " (Waugh 1968, p. 1).

This element of uncertainty is reflected in the development of the 1966 vintage French burgundies. These wines were considered years away from full ripeness when, in 1975, out went the cry, "Drink your 1966's!"; they had unexpectedly gone over the hill (Robards 1976, p. 38).

Such events explain some of the sudden downturns in wine prices. Figure 1 traces both the expected and actual maturation of a hypothetical vintage and its price through time. If, at some point in the life of the wine, investors discover that it has peaked prematurely, then they readjust their offer price to reflect this new information. On the other hand, a publicized tasting which revealed much life in a wine of old vintage could be expected to raise its value to the investor.

Quality uncertainty need not, by itself, imply that wine will command a risk premium unless the variance in the return to wine is large relative to the total variance of a portfolio. If wine speculators do not hold a "market" portfolio, such a possibility may exist.

In addition to unanticipated quality changes, wine is affected by cyclical demand fluctuations which can severely affect wine prices. For

example, the 1974–75 recession coincided with a decline in wine prices. This covariance with other asset returns is an additional reason that wine will command a risk premium.

In view of the uncertainty associated with wine's growth and development and of wine's covariance with other assets in a potential portfolio, Krasker's empirical finding that investors receive no risk premium for holding wine is surprising. This paper suggests an alternate method for estimating the wine investor's return.

My approach differs from Krasker's in three important respects. Krasker's data on wine prices are taken from the annual Heublein Wine Auction and cover the 4 years 1973–74, 1974–75, 1975–76, and 1976–77. By limiting his data to these 4 years in particular, he has biased his procedure toward finding a subnormal rate of return. Two years of his sample of 4 were extremely poor from the perspective of the wine industry; in early 1974 and through 1975, newspaper headlines proclaimed that "rivers of surplus wine" existed, reflecting the high inventories of wineries and their distributors. In these 2 years, wine prices, the French especially, fell precipitously below those prices of many preceding years. The negative trend of these figures could be expected to impact heavily on wine returns calculated during this period. My data include the prices determined by the Heublein Auction over 8 years, beginning in 1969 and ending with the 1977 auction. Observations covering 4 more years may offset the dampening effect of the 1974–75 and 1975–76 auctions on the total return to wine storage.

Second, I eliminate an unknown in Krasker's regression: the cost of storing wine. Where his regression estimate of the cost to the wine investor is $16.80 per case per year, I use the cost accounting figure of Freemark Abbey Winery, an annual per case storage cost of $.449. This number is calculated on the basis of the opportunity cost of the storage facility divided by the number of cases. The Freemark Abbey storage cost figure should reflect the cost for the auction participants who, in general, store wine in commercial cellars.

For comparison, however, I estimate the annual cellar costs and the rate of return to wine storage; in this way I have a basis for comparing Krasker's and my results.

Finally, my estimation procedure will differ from Krasker's generalized least-squares process. Limitations on available computer software forced me to employ an alternative estimation procedure which I explain below.

The Model

Following Krasker, I restrict my attention to include only red Bordeaux and California Cabernet Sauvignon produced since 1950.

Thus, I concentrate on those wines which are commonly found to benefit by extended aging, and I exclude those wines which Krasker believed were traded as antiques.[1] An observation is recorded every time a particular vintage wine is traded in consecutive years; then over the year t, $t + 1$, the rate of return to wine i may be represented by $(p_{ti+1i} - p_{ti})/p_{ti}$, where p_{ti} is the price of wine i 1 year prior to the price observation on the same wine, p_{t+1i}.

Krasker tests the hypothesis that the expected rate of return to storing wine is equal to the rate of return to riskless assets; that is,

$$\frac{Ep_{t+1i} - p_{ti} - \delta}{p_{ti}} - r = \theta = 0, \tag{1}$$

where θ is the risk premium, δ is the cost of wine storage, and r denotes the average rate of return to 3-month Treasury bills for the four quarters which precede the auction. Equation (1) implies the following regression equation:

$$\frac{p_{t+1i} - p_{ti}}{p_{ti}} - r_t = \theta + \delta \frac{1}{p_{ti}} + \frac{u_{t+1i}}{p_{ti}}, \tag{2}$$

where the error term u_{t+1i} is serially uncorrelated as expectations are assumed to be rational, and u_{t+1i}/p_{ti} is assumed to have mean zero and constant variance.

Because the disturbances are correlated within each year, Krasker employed a GLS procedure to derive the estimated equation (standard errors in parentheses):

$$\frac{p_{t+1i} - p_{ti}}{p_{ti}} - r_t = \underset{(.037)}{-.059} + \underset{(.72)}{1.4} \frac{1}{p_{ti}} + \frac{u_{t+1i}}{p_{ti}}, \tag{3}$$

$$R^2 = .03; \text{SE} = .24.$$

The estimate implies that wine costs \$1.40 per bottle (\$16.80 per case) to store. Although not significantly different from zero, the return differential between wine and Treasury bills is negative.

The Fixed-Effects Model

Because the sample data involve different numbers of observations for each different year, the GLS approach is not straightforward. Krasker does not mention how he estimates the off-diagonal component of the variance-covariance matrix for his GLS estimator, and

[1] Although prestige and publicity over bottle content assuredly guided the purchaser of the 1806 Chateau Lafite, many truly fine wines have been excluded by the arbitrary choice of 1950 for a cutoff point. Nonetheless, I strive to maintain the similarities of these two papers where I can, so I have adopted this same breakpoint.

instead of writing my own seemingly unrelated regression program, I shall consider an alternate method to account for disturbances which are correlated within each year: least-squares estimation with dummy variables. The dummy variable technique exploits the special feature of the combined cross-section/time-series data by explicitly accounting for the "constant" portion of each error term for each given cross section.

As in the GLS procedure, this method presumes that each residual u_{t+1i}/p_{ti} for observations in a given year may be decomposed into two independent components: an individual effect, v_{ti}, and a remainder, ϵ_t. Then, $u_{t+1i}/p_{ti} = \epsilon_t + v_{ti}$, where, in terms of the present example, ϵ_t represents a nonrandom addition to the rate of return to all wine sold in the year t, and v_{ti} becomes the specific classical error term associated with the ith wine, that is,

$$Ev_{ti}v'_{t'j} = \begin{cases} \sigma_v^2, \text{ for } t = t' \text{ and } i = j \\ 0, \text{ for } t \neq t' \text{ or } i \neq j. \end{cases}$$

The ϵ_t component of the rate of return can be incorporated into Krasker's regression equation by the introduction of dummy variables to represent given years. In this way, each year is characterized by its own special intercept. The regression equation is then a fixed-effects model, using Scheffé's (1959) terminology.

For ease of exposition, I will denote the dependent variable as y_{ti}. The vector $\mathbf{y}_k$ ($k = 1, \ldots, m$) will contain all observations on the dependent variables for year k. The variable $1/p_{ti}$ will be denoted z_{ti}, and $\mathbf{z}_k$ will be the vector of observations on z_{ti} for year k. The vector $\mathbf{v}_k$ ($k = 1, \ldots, m$) will contain all of the observations in v_{ti} for year k. Finally, $\mathbf{i}_k$ ($k = 2, \ldots, m$) represents a vector with entries of one in the elements associated with the kth year and zeroes elsewhere. Stacking the vectors, let

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_m \end{bmatrix}, \ \mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_m \end{bmatrix}, \text{ and } \mathbf{v} = \begin{bmatrix} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_m \end{bmatrix}.$$

The relationship to be estimated from the combination of time-series and cross-section data can then be expressed in the form

$$\mathbf{y} = \mathbf{c} + \beta_2\mathbf{i}_2 + \ldots + \beta_m\mathbf{i}_m + \delta_\mathbf{z} + \mathbf{v}, \tag{4}$$

where (c, β_2, β_3, . . . , β_m, δ) is a vector of unknown parameters. Equation (4) can be estimated efficiently with ordinary least squares. The parameter estimates can be used to estimate the quantity (c + $[1/m]\ \Sigma_{i=2}^{m}\ \beta_i$); the risk premium that one would expect to accrue to a wine is randomly selected from the total sample.

The application of this procedure to the wine data indeed reveals a premium to the wine investor with one qualification. Running four

TABLE 1

THE RETURN TO WINE STORAGE

	Regressions Using 4-Year Data (N = 140)		Regressions Using 8-Year Data (N = 199)	
	Regression Estimate of Storage Cost	Storage Cost Estimate $.449	Regression Estimate of Storage Cost	Storage Cost Estimate $.449
Average premium to wine investor	−.01367	.03812	.08542	.16614
Standard error (average premium)	.03399	.0661	.03379	.03296
t-statistic	−.4021	.5767	2.5279	5.0392
δ	17.25	...	16.60	...
(SE)	(8.76)		(5.13)	
R^2	.29	.27	.41	.38
Standard error	.2415	.2439	.2532	.2587
Mean dependent variable	.0182	.0168	.1036	.1068

separate regressions, I have distinguished between the effects on the rate of return of (1) increasing the period of observations from 4 to 8 years and (2) estimating (on the right-hand side) versus incorporating (on the left-hand side) the storage cost in the regression equation. Specifically, I run two regressions on equation (4), one using Krasker's 140 observations taken over 4 years and the second using 199 observations from an 8-year time span. In an identical fashion, I run two more regressions on the following version of equation (4),

$$\mathbf{y} - \delta\mathbf{z} = c + \beta_2\mathbf{i}_2 + \ldots + \beta_m\mathbf{i}_m + \mathbf{v}, \tag{5}$$

where δ = .449, the Freemark Abbey storage cost estimate. Using the estimates for $(\hat{c}, \hat{\beta}_2, \ldots, \hat{\beta}_m)$, I have calculated $\hat{c} + (1/m)\sum_{i=2}^{m}\hat{\beta}_i$ together with its standard error. The results appear in table 1.

Not surprisingly, one estimate confirms Krasker's findings that the expected return to wine lies beneath the return to Treasury bills. This regression utilized Krasker's data over the period 1973–77, where 1973–75 were very poor performance years for the wine industry. And like Krasker's result, the return is biased downward by the unrealistic storage cost estimate produced by the regression: an annual $17.25 per case. The influence of this estimate can be seen in the second regression, where the cellar cost is accounted for using the Freemark Abbey winery estimate; a 3.8 percent premium to oenophiles separates the wine and Treasury investment. However, as in Krasker's study, neither premium is significantly different from zero.

Expanding the data to cover the 8-year span (1969–77) greatly

improves the calculated return to wine. A low dividend of 8.5 percent reflects the high estimated cellar cost of $16.60. But wine outperforms Treasury bills by an impressive 16.6 percent for the commercial establishment for whom the Freemark Abbey cost gauge applies. Both premiums are significantly different from zero.

Reinterpreting Krasker's Regression Equation

A somewhat disturbing question which I have yet to address is why Krasker's and my regression equation with unrestricted storage costs, that is

$$\frac{p_{t+1i} - p_{ti}}{p_{ti}} - r_t = \theta + \delta \frac{1}{p_{ti}} + \frac{u_{t+1i}}{p_{ti}},$$

give the implausible estimate, attributed to $1/p_{ti}$, of $16.60 per case annually. However, note that $16.60/p_{ti}$ is a quantity on the order of magnitude of a rate of return, since the average value of p_{ti} over the 8-year period is $410. An alternative interpretation of this result is that $\delta(1/p_{ti})$ does not represent annual storage costs but rather an as yet unexplained portion of the rate of return to wine storage.

If the premium of the wine investor can be segregated into two effects, one which accrues to the wine regardless of price and a second effect related to its nominal price, then $\theta + (\delta/p_{ti})$ will represent the total return to storing wine. In this case δ/p_{ti} captures the premium as a function of price. Krasker's positive estimate for δ, interpreted in this light, implies a higher premium to lower-priced wines. The alternative postulate that high-priced wines receive the better rate of return would be supported by a negative δ.

The possibly surprising conclusion that lower-priced wines receive higher premiums than their counterparts can be explained by risk. Wines attached to smaller nominal price tags for some reason entail greater risk and return for the investor. Thus price stands as a proxy for risk, where risk can be measured by the variance of the rate of return. The validity of this explanation for Krasker's "cost" estimate relies on the variance of the rate of return of high- and low-priced wines, and support for this interpretation of the regression equation would be gained by a measurably higher rate of return variance in the low-price group. Ordering the 199 observations from the 8 auction years by nominal price, I have computed the mean rate of return and variance for the 65 highest- and 66 lowest-priced wines. While wines in the high price range offered the investor an average rate of return of .03 (3 percent) with a variance of .051, the lowest one-third provided a mean rate of return of .3587 (36 percent) with a variance

more than twice as large as the highest third, .1162. Performing an F-test, I can reject the hypothesis that the two variances are equal at the 95 percent level of significance, since the value of the F-statistic is 2.28.

Repeating the procedure using 140 observations from the 4 auction years considered by Krasker yields similar results. The 47 highest-priced wines provide a mean return of $-.0267$, with a variance of .044; the 47 lowest-range wines earned an average return of .1843 (18 percent) with a variance of .086, double that of the high-priced wines. Again, based on an F-statistic of 1.95, I can reject the hypothesis that the two variances are equal at the 95 percent level of significance.

Apparently, then, wine has an even higher return than allowed earlier in this paper, since previous calculations did not incorporate the return associated with a wine's own price but rather attributed that portion of the rate of return to cost. Using the cellar cost statistic computed by Freemark Abbey Winery, we can infer from Krasker's regression that

$$\frac{p_{t+1i} - p_{ti}}{p_{ti}} - r_{ti} - .449 \frac{1}{p_{ti}} = \left(\theta + \delta \frac{1}{p_{ti}}\right) + \frac{u_{t+1i}}{p_{ti}}. \quad (6)$$

The amended risk premium to storing wine for the 8-year period, derived by computing $\hat{c} + (1/m)\sum_{i=2}^{m} \beta + (\hat{\delta}/p)$, with $\bar{p} = \$410$, is .1239 with a standard error of .0672.

Before proposing an intuitive explanation for the greater risk associated with lower-priced wines, it should be emphasized that "lower priced" is relative to this sample of expensive premium wines. By no means do I mean to imply that risk is associated with less costly wines in general; consistency (i.e., zero risk) is the by-line of Ernest and Julio Gallo's jug wines.

One plausible interpretation of the result is that the most expensive wines, heavily represented by the older vintages, offer greater certainty to the investor because of their age. Their more youthful counterpart, with a shorter track record, provides the wine buyer with unrealized potential, and thus risk, at a discounted price.

Conclusion

In reestimating the rate of return premium to wine storage using a larger data set, I have reversed Krasker's findings that wine speculation yields no greater return than do Treasury bills. In addition, I have shown that the use of an actual storage cost measurement allows reinterpretation of Krasker's estimated "storage cost" as part of the return to wine. Adding this return to Krasker's original estimates also produces a positive premium to wine storage.

References

Goldwyn, Craig. *Chicago Tribune* (June 18, 1979).

Krasker, William S. "The Rate of Return to Storing Wines." *J.P.E.* 87, no. 6 (December 1979): 1363–67.

Robards, Terry. *The New York Times Book of Wine*. New York: New York Times Book Co., 1976.

Saturday Review of Literature. "Liquid Assets, Wine as an Investment" (July 7, 1979).

Scheffé, Henry. *The Analysis of Variance*. New York: Wiley, 1959.

Waugh, Harry. *The Changing Face of Wine: An Assessment of Some Current Vintages*. London: Wine and Spirit Publications, 1968.

On the Relationship between Commodity Price Changes and Factor Owners' Real Positions

James Cassing

Australian National University

Recent contributions to trade theory have utilized the generalized variable proportions model in order to identify "natural friends" and "natural enemies" of particular relative commodity-price changes (see, e.g., Jones and Scheinkman [1977] and Jones [1979, chap. 8] for a relevant survey). In the case where the number of factors equals the number of commodities, it is now known that every factor has a natural enemy in some commodity in the sense that any factor's reward is inversely related to some commodity's price. Also, it is known that, in general, a given factor need not have a natural friend in the sense that the factor's reward may not increase more than proportionately to any commodity's price increase.

However, this result should not be interpreted to imply that in the $n \times n$ case some individual whose endowment consists only of a given factor may not be able to increase his real income through an appropriate commodity-price rise. On the contrary, when we take into account an individual's expenditure pattern, it can be shown that any individual can improve his real position through some commodity-price rise.

In order to see this, consider the standard $n \times n$ variable proportions trade model. Let Θ denote the $n \times n$ distributive share matrix, w the $n \times 1$ vector of factor rewards, and p the $n \times 1$ vector of commodity prices. Using "ˆ" to denote a proportionate change, we have the usual price equation

$$\Theta'\hat{w} = \hat{p}. \tag{1}$$

This paper has benefited from the helpful comments of Peter Warr and an anonymous referee.

[*Journal of Political Economy*, 1981, vol. 89, no. 3]

We assume that Θ' is nonsingular so that

$$\hat{w} = \Theta'^{-1}\hat{p}. \tag{2}$$

Jones and Scheinkman show with an example that for any factor i the ith row of Θ'^{-1} may contain elements all less than unity.

But the change in an individual's real position also depends upon expenditure patterns. Suppose that agent l's endowment consists only of the ith factor. Denote by ϕ^l_{ik} the lth agent's expenditure share of income on commodity k. Then the real income effect for individual l of a change in any commodity k's price is given by

$$\hat{w}_i - \phi^l_{ik}\hat{p}_k. \tag{3}$$

Substitution from (2) yields the real income effect

$$(\Theta'^{-1}_{ik} - \phi^l_{ik})\hat{p}_k, \tag{4}$$

where Θ'^{-1}_{ik} is the $i - k$th element of Θ'^{-1}.

The real position of agent l improves if the term in parentheses in equation (4) is positive. We know that this term is negative for some k, since all factors have natural enemies. But the sum of these coefficients over all k is zero, since Θ' is row stochastic, so that Θ'^{-1} has unit row sums, and since all income is spent. That is,

$$\sum_{k=1}^{n} \Theta'^{-1}_{ik} - \sum_{k=1}^{n} \phi^l_{ik} = 1 - 1 = 0. \tag{5}$$

Therefore, since the summation has at least one negative term, it must have a positive term as well. The real position of any individual can be improved through an appropriate commodity-price rise.

As it stands, this result is not the exact counterpart of the "existence of a natural enemy" result which depends on only the technology. A natural friend would be a commodity for which a price rise induces an improvement in a factor's real wage, independent of the expenditure shares of its different owners used to compute each owner's price deflator. Rather, we show here only that in the $n \times n$ case for each individual l whose endowment consists only of a given factor, there exists a commodity such that, when this commodity price rises, the real wage of that factor, when deflated by l's price index, rises. Thus every individual is a proponent of some—not necessarily the same—commodity-price increase. And, of course, if all owners of a single factor happen to display the same expenditure shares, then there would be a sense in which every factor can improve its real position through some commodity-price rise.

References

Jones, Ronald W. *International Trade: Essays in Theory.* Amsterdam: North-Holland, 1979.

Jones, Ronald W., and Scheinkman, José A. "The Relevance of the Two-Sector Production Model in Trade Theory." *J.P.E.* 85, no. 5 (October 1977): 909–35.

Confirmations and Contradictions

Stochastic Implications of the Life Cycle–Permanent Income Hypothesis: Evidence for the U.K. Economy

Vince Daly and George Hadjimatheou

Kingston Polytechnic

In a previous issue of this *Journal,* Hall (1978) presented a simple life cycle–permanent income model of the consumption decision according to which consumption expenditures are predicted to follow a random walk with trend; that is,

$$C_t = \lambda C_{t-1} + \epsilon_t, \qquad \lambda > 1, \tag{1}$$

where ϵ_t is white noise. Hall's stated purpose in deriving this equation was to provide a testable implication of the pure life cycle–permanent income hypothesis which avoided the econometric problems associated with a consumption function in which measured income is included as an explanatory variable. The testable implication is that, apart from this period's consumption expenditure, no variable observable in this or in earlier periods should show any predictive power for next period's consumption expenditure. For U.S. quarterly data 1948(1)–1977(1), Hall found this to be the case, with the exception of movements in the real value of corporate stock.[1] Given the importance of the hypothesis it would seem worthwhile to examine the extent to which it is substantiated by alternative economic histories, and so we present here the results obtained from U.K. quarterly data in the postwar period.

[1] In Hall's paper the significance of the corporate stock variable is not considered as contradictory to the tested hypothesis when the model is modified to allow for a lagged adjustment of consumption to a change in permanent income. The implicit assumption is that, given the random walk nature of corporate stock values, changes in them are expected to be correlated with changes in permanent income.

[*Journal of Political Economy,* 1981, vol. 89, no. 3]

The data are taken from *Economic Trends—Annual Supplement 1980*, and, following Hall's example, variables were, where appropriate, measured per capita and deflated by the implicit price deflator of nondurable consumption expenditure, the latter being the measure of consumption adopted in the original study. Since the test procedure involved searching for variables which could supplement the predictive power of equation (1), a large number of regression results were generated which we do not report below; we simply note that unemployment, the inflation rate, and the rate of interest, all of which have been proposed on occasion as determinants of C_t, did not contribute significantly to the explanatory power of a regression equation containing C_{t-1}.

Our substantive result is that regression specifications which include lagged values of disposable income or liquid assets, or the period-to-period changes in these variables, and regressions which include consumption expenditures lagged by more than one period can significantly outperform the simpler specification in which consumption expenditure is predicted only by its own lagged value.[2] Thus some doubt is cast on both the validity of the pure life cycle–permanent income hypothesis, in the case of the United Kingdom, and the policy prescriptions that might follow from an acceptance of the hypothesis.

It can be seen in the Appendix that both consumption expenditure lagged more than once and lagged income appear with coefficients which are significantly different from zero. Furthermore, for the shorter period 1964–78 for which data on liquid assets are available, changes in the latter in the recent past seem to contribute significantly to the explanatory power of the regression. In all cases the size of the F_1-statistic suggests that the hypothesis that the additional variables do not contribute to the regression should be rejected.[3]

Finally, repeated estimation for different subperiods of the simple regression of C_t on C_{t-1} indicates that the coefficient on the latter

[2] The relevance of liquid assets in explaining consumers' expenditure has been suggested in a number of recent studies; see, e.g., Townend 1976.

[3] When 10 observations of each sample are retained for a postsample parameter stability test of the alternative specifications, the χ^2 (Davidson et al. 1978, p. 674) and Chow's F-test suggest that in all cases the hypotheses of postsample parameter stability and of no structural change should not be rejected. What is more important, however, for the present study is that the specification which allows for the influence of changes in liquid assets seems to provide more accurate forecasts than the pure life cycle–permanent income hypothesis. The superiority of the alternative specification is also maintained when stability is tested for post-1972 data—a critical period for econometric forecasting, coinciding with a dramatic increase in the U.K. saving ratio (Hadjimatheou 1979). For this period, however, the value of χ^2 suggests that the hypothesis of postsample parameter stability should be rejected for both specifications while the F-value is still not statistically significant.

variable is unstable. More specifically, as can be seen from the following values of $\hat{\lambda}$, the greater the time span of the sample, moving backward toward the less recent past, the higher the level of the coefficient: (i) 1973(1)–1978(4), 1.00125; (ii) 1964(1)–1978(4), 1.00374; (iii) 1956(1)–1978(4), 1.00413; (iv) 1964(1)–1972(4), 1.00483; (v) 1956(1)–1972(4), 1.00497 (Chow test: iii and v, $F(24,66) = 3.14$; ii and iv, $F(24,34) = 2.21$). In the context of a more conventional consumption function this is not surprising, given the marked downward trend of the average propensity to consume in the United Kingdom, especially in the 1970s. This observation, however, does not seem to lend support to Hall's statement that constancy of λ is expected to "be a good approximation, at least over a decade or two" (Hall 1978, p. 975).

In conclusion, the present evidence seems to suggest rejection of the pure life cycle–permanent income hypothesis in the U.K. case. Furthermore, the latter cannot explain the revealed differences between the two countries. It may be tempting to suggest that English people are less "rational" than Americans, but such a pronouncement would certainly need an explanation.

What is probably needed is a more general theory of the consumption decision which encompasses the different experience of the two countries as special cases. It may be that binding financial constraints, changes in the age structure and the distribution of income, as well as changes in interest rates and the composition of wealth, are too important to be ignored. The importance of some of these factors has, after all, been recognized, either explicitly or implicitly, by the life cycle and/or permanent income hypotheses.

Appendix

Dependent Variable is C

1956(1)–1978(4), $N = 91$:

$$\underset{(1055)}{1.00413C_{t-1}}, \tag{A1}$$

$$\bar{R}^2 = 0.994, \quad S = 2.103, \quad SSR = 398.2.$$

$$\underset{(7.9)}{0.819C_{t-1}} + \underset{(2.1)}{0.2197C_{t-2}} + \underset{(2.9)}{0.1201Y_{t-1}} - \underset{(3.6)}{0.1492Y_{t-2}}, \tag{A1.1}$$

$$\bar{R}^2 = 0.995, \quad S = 1.986, \quad SSR = 343.1, \quad F_1(3,87) = 4.66.$$

1964(1)–1978(4), $N = 59$:

$$\underset{(780)}{1.00374C_{t-1}} \tag{A2}$$

$$\bar{R}^2 = 0.971, \qquad S = 2.441, \qquad SSR = 346.0.$$

$$\begin{aligned} &\underset{(4.1)}{0.4971C_{t-1}} + \underset{(1.67)}{0.2353C_{t-2}} + \underset{(2.3)}{0.273C_{t-3}} \\ &+ \underset{(1.92)}{0.0895\Delta Y_{t-1}} + \underset{(5.2)}{0.115\Delta LA_{t-1}}, \end{aligned} \tag{A2.1}$$

$$\bar{R}^2 = 0.9884, \quad S = 1.881, \quad SSR = 191.0, \quad F_1(4,54) = 10.96.$$

$$\begin{aligned} &\underset{(4.8)}{0.604C_{t-1}} + \underset{(3.6)}{0.442C_{t-2}} + \underset{(1.60)}{0.0831Y_{t-1}} \\ &- \underset{(1.79)}{0.105Y_{t-2}} + \underset{(3.8)}{0.0886LA_{t-1}} - \underset{(4.0)}{0.0926LA_{t-2}}, \end{aligned} \tag{A2.2}$$

$$\bar{R}^2 = 0.988, \quad S = 1.96, \quad SSR = 204, \quad F_1(5,53) = 7.38.$$

Numbers in parentheses are t-statistics; S = standard error of estimate, SSR = sum of the squares of the residuals, and F_1 is a test for the significance of additional parameters; $\Delta Y_{t-1} = Y_{t-1} - Y_{t-2}$ and $\Delta LA_{t-1} = LA_{t-1} - LA_{t-2}$; LA = selected liquid assets of the personal sector as given in *Financial Statistics*.

References

Davidson, James E. H.; Hendry, David F.; Srba, Frank; and Yeo, Stephen. "Econometric Modelling of the Aggregate Time-Series Relationship between Consumers' Expenditure and Income in the United Kingdom." *Econ. J.* 88 (December 1978): 661–92.

Hadjimatheou, George. "An Explanation of the Decline in the Average Propensity to Consume in the U.K." *British Rev. Econ. Issues* 2 (November 1979): 1–13.

Hall, Robert E. "Stochastic Implications of the Life Cycle–Permanent Income Hypothesis: Theory and Evidence." *J.P.E.* 86, no. 6 (December 1978): 971–87.

Townend, J. C. "The Personal Saving Ratio." *Bank of England Q. Bull.* 16 (March 1976): 53–73.

Miscellany

Why There Are No Risk Preferrers

Somewhat more than 100 years ago, Smith's celebrated diamond-water paradox was solved by the observation that however valuable water might be to someone who had none, if it was produced and sold in unlimited quantities at a price p consumers would adjust their rate of consumption so that its marginal value to them was at most p. The same argument implies that behavior under uncertainty depends not merely on the actor's utility function but also on the cost of producing (and absorbing) risk. However fond I may be of uncertainty, I will not buy it at a high price (by choosing a profession whose returns have high variance but low expected value) when I can get it at a low one (in Las Vegas).

Risk can be produced inexpensively; existing sellers in Las Vegas and elsewhere charge only a few percent of the expected value of the gamble, part of which must be regarded as payment not for the risk itself but for the palatial surroundings in which it is produced. Hence, absent special circumstances (such as a government monopoly of risk production), we may expect individuals in equilibrium to exhibit at most a small degree of risk preference, for the same reason that they exhibit at most a small marginal utility for water.

Insurance, unlike risk, is inherently costly to produce because of the problems of moral hazard. Moral hazard is a major cost of insurance but a minor cost of gambling because a gambling game, being created for the purpose of producing risk, can be designed to almost eliminate the ability of participants to spend resources on influencing the outcome. The seller of insurance faces a preexisting situation in which it is likely that the buyer (of, say, unemployment insurance) can and will affect the odds by his actions. Hence insurance is costly to produce and individuals in equilibrium may exhibit a substantial degree of risk avoidance.

David Friedman

Virginia Polytechnic Institute

[*Journal of Political Economy*, 1981, vol. 89, no. 3]

Book Reviews

The Regulatory Process and Labor Earnings. By RONALD G. EHRENBERG. New York: Academic Press, 1979. Pp. xx+204. $15.00.

The irony is that this study was not published in the *Bell Journal of Economics.* Ehrenberg's argument is that the wage rates of workers in the New York Telephone Company are above those of other equally qualified workers and that therefore the New York Public Service Commission should not allow further wage increases to be passed on as higher telephone rates. As I read through his argument I wondered how Ehrenberg would have treated a request to cover increased publication costs by the editors of the *Bell Journal.* A related tactic did not escape notice by the leaders of the Communication Workers of America (CWA), however, who were on the other side of the regulatory proceeding in which Ehrenberg participated. In addition to implying that Ehrenberg was a mere hired gun for Alfred Kahn (then chairman of the New York State Public Service Commission), according to Ehrenberg, ". . . political pressure was put on my employer, the New York State School of Industrial and Labor Relations (ILR) at Cornell University, by the New York State AFL-CIO, which cancelled a scheduled conference at Cornell, implicitly suggesting that it would be wise if ILR faculty did not get involved in similar cases in the future" (pp. 147–48). The hero of this part of the story was apparently Bob McKersie (then dean of the ILR school but now at MIT) who refused to take institutional responsibility for an individual faculty member's research conclusions.

I think most economists will find this a fascinating study. In fact, I would urge any economist looking for a good read while on holiday to take this (relatively inexpensive) book along. I took it and a pretty good paperback for a week on Martha's Vineyard (spending a whole summer there is apparently a prerequisite for becoming president of the AEA) and actually finished Ehrenberg's book first. For one thing, it is well written. Regression equations and the like are kept in an appendix. As a consequence the book would make excellent supplementary reading for graduate or undergraduate courses in labor economics, industrial organization, public policy, or applied econometrics.

For another thing, the book tells a coherent story about a complete policy-oriented research project from start to finish. The 1976 inception of Ehrenberg's study resulted from an unprecedented request by the New York Telephone Company for a rate increase to cover future, yet unnegotiated,

[*Journal of Political Economy*, 1981, vol. 89, no. 3]

increases in wages and benefits. The conclusion of the study comes a year later when the Public Service Commission declines to accept Ehrenberg's logic and therefore finds his empirical calculations inappropriate to the decision at hand. In the meantime, however, the union-management negotiations had been completed under the shadow of a possibly adverse decision on cost pass-through, and, I would bet, this was just the purpose for which Ehrenberg's study was initiated by the Public Service Commission staff in the first place.

Ehrenberg's study is also coherent in another dimension. It reflects a remarkable breadth of substantive argument over a range of subjects covering labor law, the determinants of unionization and bargaining structure, as well as some very convincing empirical economics. For labor economists there are some special treats. Ehrenberg tells us that he billed the Public Service Commission staff for 40 days of consulting time (at an undisclosed rate), while National Economic Research Associates billed New York Telephone for 87 days of consulting time (at an undisclosed, but higher, rate). I am not sure what this is supposed to imply for public policy, but it does raise a larger issue. As the real academic incomes of economists have fallen throughout the last decade (I recently calculated that junior faculty taking jobs at my university this year will receive real salaries 22 percent below the starting salaries 12 years ago), their efforts have naturally turned toward the more practical and remunerative. One cannot help wondering what effect this trend will have on the profession's research output as this income squeeze intensifies during the decade of falling college enrollments that we are entering.

The core of Ehrenberg's study is his empirical comparison of wage rates and fringe benefits at New York Telephone with the wage rates of other workers. Ehrenberg deliberately chooses "other workers" to be other nonunion workers with similar skills, a decision the Public Service Commission ultimately found inappropriate. Wage comparisons are first made using several surveys from the Bureau of Labor Statistics and other sources. Since worker human capital is difficult to control in these comparisons, standard wage regressions from the 1/100 sample of the 1970 Census data for New York State are also computed. The telephone company is so large an employer, and New York so large a state, that this procedure produces relatively precise wage-differential estimates. Taken together, I found Ehrenberg's empirical case convincing, especially when bolstered with the extraordinarily low relative quit rates at New York Telephone. In sum, I think there is little doubt that wage rates at New York Telephone had by 1976 been increased above market-clearing levels and that many of the employees of this company were earning economic rents. The question that naturally arises, therefore, is what attitude the New York Public Service Commission should have taken toward this stylized fact on the occasion of a rate-increase request.

The administrative law judge in this regulatory proceeding recommended that this "fact" be ignored because to do otherwise would be to interfere in the collective-bargaining process in violation of federal labor law. Thus, the administrative law judge concluded that whether to use Ehrenberg's wage comparisons was mainly a "legal issue." As anyone who has consulted with lawyers will understand, a "legal issue" is primarily one for which there is no accepted answer. In fact, Ehrenberg shows that as a matter of law the administrative law judge was simply incorrect on this point and the Public Service Commission concluded similarly. Despite the commission's conclusion that disallowing excessive labor-cost increases was legally permissible, however, the commission explicitly rejected Ehrenberg's wage comparisons as inappropriate.

In effect, Ehrenberg compared the average wage rate in a unionized monopoly firm against the average of all other wage rates, union and nonunion. Since the majority of all workers are nonunion, this is primarily a comparison of the wage rates of unionized workers in a monopoly firm against the wage rates of nonunion workers. The commission argued that the average wage rate of the workers at New York Telephone should have been compared against the average wage rate of other unionized workers. Since there is an empirically established overall union/nonunion wage differential, this comparison would essentially have uncovered whether the union/nonunion wage differential was systematically greater in a particular regulated-monopoly industry. If this had been the case, the commission apparently would have been prepared to disallow further labor-cost increases as within the scope of its mandate to regulate monopoly power but not union power.

Ehrenberg really offers no convincing argument against the commission's position. Moreover, although he does not offer the necessary empirical work to test the appropriate factual issues, I suspect that such a study would not demonstrate that union/nonunion wage differentials were higher among unionized employees of monopoly firms than among other unionized workers. Historically, for example, union/nonunion wage differentials have been highest among construction workers and mine workers, both relatively competitive industries.

The difficulty here is that the commission's position essentially takes the presence of unionization as exogenous and asks merely whether the presence of regulated monopoly exacerbates union wage pressures. This position ignores the really important question of whether an industry might not have become unionized *because* it is a regulated monopoly. In this case the presence of the union is endogenous to the regulatory process, and the appropriate wage comparison is akin to the one Ehrenberg offers. The anecdotal evidence he presents of the evolution of the Bell System bargaining structure as well as the well-known anecdotal evidence regarding the trucking and railroad industries and the history of the Davis-Bacon Act in contract construction all support this interpretation. As Ehrenberg's experience clearly demonstrates, however, making this case before a regulatory commission is likely to be a thankless task, and perhaps not very remunerative either.

ORLEY ASHENFELTER

Princeton University

Law, Legislation and Liberty: A New Statement of the Liberal Principles of Justice and Political Economy. Vol. 3: *The Political Order of a Free People.* By FRIEDRICH A. HAYEK.

Chicago: University of Chicago Press, 1979. Pp. 243. $14.00; London and Henley: Routledge & Kegan Paul, 1979. Pp. 243. £5.95.

> But to avoid this [destroying our civilization] we must shed the illusion that we can deliberately create the future of mankind. This is the final conclusion of the forty years which I have devoted. . . .

I want to thank Kaj Areskoug, Fritz Machlup, Ingo Walter, and Lee Wohlfert for comments on an earlier draft of this review.

These dramatic words end the third and last volume of *Law, Legislation and Liberty: A New Statement of the Liberal Principles of Justice and Political Economy* (hereafter cited as *LLL*) by Nobel laureate Friedrich A. Hayek. The recently published volume, *The Political Order of a Free People,* analyzes the undermining of democratic principles and liberty in parliamentary democracies and concludes with an outline of a new constitutional framework.

The message of Hayek's voluminous work since World War II[1] is, to put it briefly, that the current political system of Western democracies is bound to cause increasing coercion of individuals by governments, private groups, and firms, and that governmental decisions will increasingly be contrary to the opinions of the majority. He argues in volume 3 that the unlimited powers (sovereignty) of governments—combined with the belief and expectation that governments can direct production, consumption, and the distribution of income to maximize social welfare—are the necessary and sufficient conditions for this development. The constitutional proposal is designed to subject the government to substantial constraints in its use of coercive powers.

The previous two volumes of *LLL* (Hayek 1973, 1976*b*) provide the foundation for the analyses of the political process in volume 3, but the last can be read independently. *Rules and Order* analyzes the role of rules and law in a "free" society—one in which no individual can be coerced to act in accordance with the will of others. Law must represent general principles of just conduct on which there is common agreement. Coercion must be used only to enforce such general principles applicable to all individuals. *The Mirage of Social Justice* discusses the limits of reason and rationality in social design and the impossibility of defining social justice. Some arguments in these two volumes will be referred to below to clarify the discussion of volume 3.

The "Bargaining" Democracy and Logrolling

Hayek uses the concept of the bargaining democracy to describe a system of representative assemblies within which different interest groups can trade in vote support for each other's proposals. In the theory of public finance, this is called logrolling[2] and is generally regarded as an improvement in the democratic process because the strength of preferences can be allowed to affect the outcome. Hayek, on the other hand, argues that the trade in vote support contributes to results that do not reflect the opinions of the majority. These views are clearly contradictory, yet the abstract nature of Hayek's analysis of the democratic process makes a comparison with conventional public finance difficult. However, Hayek's criticism of trade in vote support within representative assemblies is directed at the wide range of issues on which such practices are possible rather than at the principle of logrolling itself. The argument is derived from the distinction between votes on general rules of just conduct applicable to *all* citizens (i.e., *legislation* proper, such as private and criminal law) and votes on the distribution of payments and benefits among individuals and groups in society. The latter include all votes on taxes and transfer payments, allocation decisions with income redistribution implications, and "directives" that discriminate in favor of certain groups. The representative assemblies in all democracies do in fact practice logrolling on both types of issues. It is, therefore, possible that the distribution of benefits

[1] This work includes such classics as *The Road to Serfdom* (1944) and *The Constitution of Liberty* (1960).

[2] See, e.g., Tullock 1969.

on which a majority decide at the same time violates general rules on which the same majority agree. For example, a principle of taxation can be accepted while special benefits (e.g., tariffs and subsidies) with implicit positive or negative tax equivalents are passed by the same assembly, violating the agreed-on principle. It appears that the theory of logrolling is valid for votes on the distribution of benefits and payments *only* when no general agreed-on principle is applicable. Only pure collective goods are likely to satisfy this condition when a principle of taxation has been laid down, because individuals can be assigned tax equivalents for other publicly provided services and benefits.

The coercive powers of governments to benefit some groups at the expense of others, in addition to enforcing generally agreed-on principles, is the source of what Hayek calls "the miscarriage of the democratic ideal" (p. 98). These powers make it possible and advantageous to trade in specific benefits as well as to trade in benefits for support of general principles. Moreover, well-organized interests are likely to gain disproportionate weights in the democratic process—"there is no limit to the blackmail to which governments will be subject" when "no superior judiciary authority can prevent the legislature from granting privileges to particular groups" (p. 11).[3] The final outcome of the democratic process need not, then, correspond to anybody's opinion on what is right or to the will of the majority.[4] Hayek argues, in fact, that this concept *can* be stated only in terms of general rules and principles.[5] Behavioral incentives of the rules are lost and the will of the majority cannot be given contents when representatives decide *both* on general rules and on adjustments to the outcome of individuals' activities that are performed subject to these rules.

A Model Constitution

Hayek outlines a constitution designed both to make the outcome of the democratic process coincide with principles held by the majority and to minimize the degree of coercion in society.[6] The proposal is very simple: Distinguish between legislation proper and decisions on government spending, administration, and regulation by separating the two functions into two distinct assemblies elected by entirely different procedures. Make the government body subject to the general rules of conduct decided upon by the

[3] A suggestion like Friedman and Friedman's (1980) to legally prohibit the use of tariffs for the benefit of special-interest groups can obviously not get to the root of the problem of the democracy in Hayek's view as long as the government has the coercive power to use a multitude of alternative means of creating special benefits for the same groups.

[4] Note that Hayek regards the state of democracy in the United States as threatened as are the parliamentary democracies in Europe because the judiciary branch of the American government does generally not subject decisions of Congress to a test against general principles—though this may have been the intention of the Founding Fathers.

[5] This argument is developed in vol. 2 of *LLL*.

[6] Fritz Machlup has pointed out to me that one could argue that Hayek's proposal for constitutional reform contradicts his own views on the creation of human institutions. While Hayek argues that human institutions cannot be designed rationally and deliberately for the benefit of mankind, he also argues that they cannot develop spontaneously in accordance with people's preferences unless the conditions for such developments exist. The proposal constitutes a deliberate design for the purpose of creating such conditions.

legislative assembly. Hayek's model constitution also contains a general declaration of rights and an important definition of what qualifies as law—a general rule of just conduct. A constitutional court would be established to test the appropriateness of the legislature's decisions against such a definition.

The *legislative assembly* would be responsible for the body of criminal and private law, the principles of taxation, general regulations for safety and health, rules to secure competitive markets, corporate law, and the like. The coercive powers of governments would be limited to the enforcement of these general rules and principles.

The *government assembly* would decide on the use of material and human resources entrusted to the public sector. The size and the general purposes of expenditures would be limited only indirectly by the general rules of conduct set down by the legislative assembly and by people's willingness to pay taxes. The general principles of taxation decided upon in the legislative assembly would make citizens aware of their share in payments for specific services. This would prevent the current practice of disguising tax burdens to "make those who will ultimately have to bear it [the burden] as little aware of it as possible" (p. 127).

Would the proposed constitution achieve its purposes? A critical issue seems to be whether the legislative assembly can be prevented from instituting laws favoring large groups or wealthy interests. Hayek presents a number of suggestions in this regard. More important, however, is the definition of "law" and the role of the constitutional court in evaluating the constitutionality of legislation against this definition. Hayek characterizes (p. 109) a law as negative in the sense that it does not "aim at achieving concrete purposes," it must be applicable to an "indefinite number of unknown future instances," and it must exclude "all provisions intended or known to affect principally particular identifiable individuals or groups." This definition would apparently rule out legislation covering such issues as minimum wages and price controls. However, the boundary between measures to improve information spreading in the market and the protection of particular firms or groups is sometimes extremely vague, as in the case of legislation on health standards, consumer protection, and occupational safety. It is therefore easy to imagine situations in which the members of the constitutional court assume a critical role. The experience with the U.S. Supreme Court indicates that the existence of a constitutional court does not necessarily provide a complete safeguard against legislation that favors large groups in society. A simple addition to Hayek's proposal could help achieving his purpose—Wicksell's old suggestion (Wicksell 1896) of qualified majorities and relative unanimity. A principle that legislation must be passed by, for example, 90 percent of the assembly hardly seems objectionable since decisions on general rules must reflect the standards of most citizens to be respected and justify coercion in enforcement.

Where does Hayek's proposal lead? Debate on it cannot be carried out without reference to more specific results. Since the author provides only general and abstract clues, I will here attempt to illustrate some potential consequences by relating it to a few economic political issues.

The inflationary bias of current democracies has been a theme in Hayek's earlier writings and is clearly linked to the subject of this volume since inflation is one way for governments to disguise the true tax burden. The independence of the central bank from the government seems guaranteed under the proposed constitution because directives covering the central bank will have to take the form of general rules unrelated to specific political and economic circumstances. This provides a check on the monetization of gov-

ernment fiscal deficits.[7] Inflationary pressures could also be reduced under the proposed constitution because the government would be unable to bear the costs of unemployment and business failures. Most market interventions in the form of subsidies to labor or capital are likely to violate the general rules by which the government must abide. This could contribute to higher wage and price flexibility.[8] Furthermore, governments must consider its expenditures more carefully when the corresponding tax burden for each individual has been established by the legislature.[9]

A more controversial consequence of the proposed constitution would be that developments in production, consumption, and income distribution to which a majority object have to be accepted unless a general rule has in fact been violated. However, the outcome of economic activities may shed new light on a particular principle and lead to a change therein. The extremely far-reaching implications of this can be illustrated with the examples of immigration laws and equity-oriented policies.

Immigration restrictions appear to be unconstitutional, as do tariffs or import quotas, under Hayek's proposal. Such policy measures are clearly directed at identifiable individuals to achieve concrete purposes—economic benefits for national residents and firms. Opening up the Western industrialized economies to unlimited immigration would lead to a social transformation of enormous proportions. Most people would probably find the outcome unacceptable. Nevertheless, they would have to accept immigration as the "will of the majority" unless a general principle could be found to prevent immigration without thwarting other desirable outcomes, such as imports of competitive foreign goods.

Most policies aimed at securing certain levels of incomes for individuals or groups also have to be given up by governments that must abide by general principles. Discrimination is necessary to guarantee the outcome of economic activities pertaining to an individual. This is illustrated in volume 2 of *LLL* by a revealing and sharply satirical analysis of the contents of the United Nations *Universal Declaration of Human Rights.* This declaration contains positive rights (to outcomes) as well as traditional civil rights (to equal treatment before the law as defined by general rules of conduct). The two kinds of rights are simply contradictory. Either governments treat unequal people equally, abiding by generally accepted rules, or they treat unequal people unequally to secure certain outcomes. The only available policy instrument aimed at securing a certain income distribution would be the general tax structure. Thus, a negative income tax would have to be substituted for large parts of the social security system.

The examples above suffice to show that the constitutional proposal is far from politically feasible in the foreseeable future. There is probably no significant political group in any democracy that bases its political targets on

[7] Hayek earlier suggested the denationalization of the right to issue money (Hayek 1976*a*). Though a possibility, this suggestion is not part of the proposal here.

[8] "Social responsibility" must be the major check on wage and cost increases in an economy in which the government guarantees employment independent of the wage/cost structure (cf. Wihlborg 1978).

[9] Hayek's proposal can here be compared with the current discussion about a constitutional amendment prohibiting an unbalanced government budget. Just as tariffs cannot reduce the political pressures to discriminate in favor of a particular industry, a constitutional amendment to balance the budget does not reduce the political incentive to tax via inflation. Thus, the political process can be used, e.g., to manipulate the definition of the government budget—rendering the amendment meaningless.

general principles rather than on desired outcomes. Moreover, the proposal does not address the question of whether *any* outcome under a certain principle is acceptable. For example, is it not necessary to discriminate in order to secure survival of an individual who spends the negative income tax payments almost on receipt? Should hospital care be denied individuals who have chosen not to invest in medical insurance? Hayek has previously suggested ways of resolving these questions with a minimum of government coercion (Hayek 1960). It is not clear, however, that these solutions are commensurate with strict adherence to the constitutional principles outlined here. Despite such questions, Hayek's proposal deserves serious consideration as the starting point for further work by economists and social philosophers. A fundamental constitutional change may indeed be the only route to prevent further erosion of democratic values and individual liberty. Hayek has argued—convincingly, in my view—that the coercive powers given to governments, in the belief that progress can and should be planned and directed, are the cause of a self-generating process leading to the destruction of the decentralized market economies in which a high degree of individual liberty remains feasible.

Hayek's constitutional proposal implies quite simply that we give up the idea of steering and planning the future direction of social and economic activities. Instead, we should ask how we can set up a system of general principles under which mankind's knowledge can ensure progress for the maximum benefit of all. In Hayek's words: "To pretend to know the desirable direction of progress seems to me to be the extreme of hubris. . . . All we can do is to create favorable conditions for it [progress] and then hope for the best" (p. 169).

Clas Wihlborg

New York University

References

Friedman, Milton, and Friedman, Rose. *Free to Choose: A Personal Statement.* New York: Harcourt Brace Jovanovich, 1980.

Hayek, Friedrich A. *The Road to Serfdom.* Chicago: Univ. Chicago Press, 1944.

———. *The Constitution of Liberty.* Chicago: Univ. Chicago Press, 1960.

———. *Law, Legislation and Liberty: A New Statement of the Liberal Principles of Justice and Political Economy.* Vol. 1, *Rules and Order.* Chicago: Univ. Chicago Press, 1973.

———. *The Denationalization of Money: An Analysis of the Theory and Practice of Concurrent Currencies.* London: Inst. Econ. Affairs, 1976. (*a*)

———. *Law, Legislation and Liberty: A New Statement of the Liberal Principles of Justice and Political Economy.* Vol. 2, *The Mirage of Social Justice.* Chicago: Univ. Chicago Press, 1976. (*b*)

Tullock, Gordon. "Problems of Majority Voting." In *Readings in Welfare Economics*, edited by the American Economic Association. Homewood, Ill.: Irwin, 1969.

Wicksell, Knut. *Finanztheoretische Untersuchungen nebst Darstellung und Kritik des Steuerwesens Schwedens.* Jena: Fischer, 1896. English translation by Richard A. Musgrave and Alan T. Peacock, eds., *Classics in the Theory of Public Finance.* London: Macmillan, 1958.

Wihlborg, Clas. "Liberty and Labor Markets: A Reflection on the Swedish Experience." *ACES Bull.* 20 (Fall/Winter 1978): 85–100.

Studies in the Economics of Search. Edited by STEVEN A. LIPPMAN and JOHN J. MCCALL.
Amsterdam: North-Holland Publishing Co., 1979. Pp. 225. $41.50 (cloth).

As a sequel to their excellent survey of the job search literature, Lippman and McCall have organized an impressive collection of new papers that materially enrich the subject. The collection includes original contributions by nine young authors on topics that range from the estimation of the elasticity of unemployment duration with respect to unemployment compensation to a theoretical demonstration that searching for a bargain is the real world substitute for Walras's auctioneer. The individual essays are of uniformly high quality as are the editing and production of the volume.

According to the editors, each paper is motivated by a desire either to test the empirical implications of the so-called "standard" search model or to extend its structure for the purpose of deriving a richer menu of behavioral implications. Regarding these motives as noble, I have attempted to point out the sense in which and the extent to which each author has been led by them. Since the essays are related to the standard model in the inverse order of their presentation in the volume, my comments read from back to front.

The standard model in a labor-market context supposes that an individual worker samples wage rates sequentially from an urn of offers with a known frequency distribution at a rate of one per period. The worker wishes to maximize expected lifetime earnings net of the out-of-pocket and opportunity costs of search. As is well known, the optimal strategy is to accept the first offer in excess of some reservation wage. The reservation wage equates the cost of taking one more sample to the expected gain attributable to another sample. Under reasonable conditions, the reservation wage is independent of search duration to date.

That a reduction in the cost of search yields an increase in the reservation wage is a straightforward implication. Given the standard model, an increase in the reservation wage, in turn, implies an increase in the expected time required to find an acceptable wage and an increase in expected earnings. Because unemployment compensation must be forgone when an offer is accepted, it acts as a subsidy to search in the standard model.

Kathleen Classen tests for these predicted effects by estimating the reduced-form relationship between both unemployment duration and post-unemployment earning and unemployment compensation. Both the weekly benefit received and the maximum duration of benefits are included as right-hand-side variables. The analysis yields a strong positive relationship between duration and both measures of unemployment compensation liberality but no statistically significant relationship between eventual earning and liberality. Although the second result can be explained away, it is also true that alternative explanations exist for a positive effect on duration. The paper has many interesting things to say, but as a test of search theory, the results are inconclusive.

One would prefer to test directly for the implied relationship between the reservation wage and the cost of search. A major problem in doing so is the nonobservability of the reservation-wage rate. Nick Kiefer and George Neumann report on an attempt to get around this problem. Their method takes advantage of the structure of the standard model. In the paper, they show that data on accepted wage rates and realized unemployment durations can be used to estimate simultaneously the structural parameters of both the reservation-wage equation and the probability of finding acceptable employ-

ment function. Although their estimate of the unemployment benefit effect is positive in the first case and negative in the second as expected, the estimated effects of maximum benefit duration have the "wrong" signs. In spite of this ambiguity, the paper represents a valuable contribution to the methodology of estimating sequential, discrete choice models.

The standard model "explains" unemployment as a rational response to wage dispersion. Critics argue that this explanation is of little importance, because workers with long unemployment spells, those that account for most unemployment, do not turn down offers. They simply cannot find a job. By allowing unemployed workers to allocate time between search effort and nonmarket activity, Ken Burdett has extended the relevance of the search theoretic framework even to the case in which there is no wage dispersion. In his model, a reduction in job availability on unemployment is amplified by the rational allocation of a worker's time.

In the paper presented in the volume, Burdett shows that search intensity increases, and the reservation wage falls with unemployment duration, as the unemployment-benefit period is exhausted. This theoretical result suggests the existence of misspecification bias in both of the empirical papers discussed above. The allocation of time among search, work, and leisure by an employed worker is also investigated. This case is a natural synthesis of classical labor supply and search theory. The principal contribution of the paper is to incorporate search theoretic considerations into a more general dynamic theory of labor supply.

The papers by Jeffrey Hall, Steven Lippman, and John McCall and by John Danforth study the consequences of relaxing the assumption of risk neutrality. Presuming decreasing absolute risk aversion, the implied relationship between the reservation wage and nonhuman wealth is positive. Hence, the papers represent a rigorous derivation of the long-standing contention that the running down of financial assets explains the observed negative association between the reservation wage and duration of unemployment. As a technical matter, Hall et al. show that the reservation-wage property need not hold when past offers can be recalled. Nevertheless, the probability per period of finding acceptable employment decreases with wealth in this case as well. Finally, Danforth's formulation represents a synthesis of the theory of household wealth accumulation and the theory of wage search. As Danforth points out, the theory implies a process of increasing wealth dispersion over time if the assumption of decreasing absolute risk aversion is empirically valid. Hence, the paper is also a contribution to the theory of income distribution.

The standard search model characterizes the optimal strategy of a price-taking agent seeking a bargain. The market effect of such behavior is a reallocation of goods and services toward demanders who value them most highly. One might expect that such a process would eventually equalize marginal valuations, that is, converge to a competitive equilibrium. Joseph Wharton formalizes this idea in the context of a market for homogeneous labor services composed of workers who search on the job for a higher wage and employers who adjust wage offers in response to search-induced changes in employment. The process is modeled as a Markov chain. When the cost of search is "small," a competitive allocation of workers across firms is the only absorbing state. Although the analysis is insightful and the approach developed is amenable to further development, the employer's wage-setting behavior is not shown to be rational in any sense. Other work on this problem suggests that strategic behavior on the part of the price-setting agents can thwart convergence to competitive equilibrium.

Search-induced convergence to a single price raises the following paradox. How can the standard search model represent an important contribution to positive economics if the market effect of the behavior explained eliminates the price dispersion required to observe the behavior? Roger Kormendi provides a resolution of this paradox by formulating a market model in a pure-exchange context that exhibits price dispersion in equilibrium. One of the original features of his model is that every agent plays a role in the price formation process. The demonstration that mutually consistent (Nash) price-setting strategies exist which define a market equilibrium is a major contribution of the paper. Although a detailed discussion of the kinds of conditions one needs for equilibrium price dispersion in Kormendi's model is beyond the scope of this review, it is important to note that nonstorability of the two commodities traded is implicit. If this assumption were relaxed, Kormendi's equilibrium is only "temporary" in the sense that, once allocations have changed in response to trade, the agents have incentives to play the game again. One suspects that the repeated game might well converge eventually to a competitive single price equilibrium.

It is curious that the matching of job and worker is not explicitly recognized in the standard formulation of the individual's problem of finding acceptable employment. Louis Wilde contributes to this deficiency by distinguishing between two job characteristics: the wage, observable on initial contact, and an attribute that must be experienced on the job. The analysis characterizes an individual worker's acceptance strategy assuming knowledge of the joint probability distribution over jobs. The result is a formalization of the "job-shopping" hypothesis typically offered as an explanation for high quit rates among the young. A useful extension of this model would also recognize the employer's problem of learning about a worker's productivity.

In the final paper, Charles Stuart applies the standard search model to an entirely different subject, the spatial clustering of sellers of both the same and different commodities. Specifically, he shows that market centers are more attractive to buyers than spatially isolated sellers, because the cost of search per unit is smaller and the perceived heterogeneity of the goods offered is larger. These sources of agglomeration economies offset the effects of transportation costs and competition that would otherwise induce dispersion in the location of sales outlets. Although there are many insightful observations on the subject, the author does not attempt to derive an equilibrium solution to the location problem in which all of these forces balance. This would seem to be a fruitful area for further research.

In sum, there is much of value in the volume both to students of labor economics and to those interested in applying search theory in other areas of economic inquiry.

DALE T. MORTENSEN

Northwestern University

A Theory of Income Distribution. By HAROLD LYDALL.
New York: Oxford University Press, 1979. Pp. vii+326. $32.50.

Professor Lydall, of South African birth, received his higher education at Oxford. He has taught primarily in Britain and Australia, worked in international agencies in Geneva, and also visited extensively in the United States.

His methodology recalls the Burns and Mitchell pragmatic tradition of quantitative institutionalism. That is to say, Lydall begins by searching out statistical regularities in his data and then seeks one or more theories to explain them, preferably on more than an ad hoc basis. He looks with some suspicion on the opposite procedure of deducing a general theory first and then seeking out data to test it.

Although the present volume is indeed more nearly theoretical and less exclusively statistical than Lydall's well-known *Structure of Earnings* (1968), its title may seem misleading to the conventional economist. I myself expected a new "high theory" (Shackle's phrase), or at least a tasteful ragout of existing high theories. But after seven "background" chapters of a doctrinal-historical nature (which explain why the author feels it desirable to strike out along new paths), Lydall's independent contribution retreats more than half way to empiricism.

Once inoculated against anticipations of high theory—wage funds, iron laws, marginal productivities, widows' cruses, degrees of monopoly—the reader is impressed by Lydall's statistical generalizations, both in themselves and as challenges for integration into the rival high theories of our rival conventional wisdoms.

Lydall finds, for example, the following relation between value added per worker, which he calls v, and the average size of the employing firm as measured by employment, L: $v = BL^{\beta-1}$ ($B > 0$, $\beta > 1$). A similar relation holds for the average wage, w: $w = AL^{\alpha-1}$ ($A > 0$, $1 < \alpha < \beta$), and for one-man firms (OMF), $A = B$. The size distribution of firms follows the Pareto law, with N as the number of firms: $N = RL^{-\rho}$ ($R > 0, \rho > 1$). (In these three equations, both capital and Greek letters are statistical parameters.)

An ingenious algebraic development (pp. 297, 298) combines these results into a theoretical labor share S_w of $(\rho - \beta)/(\rho - \alpha)$, derived under conditions Lydall calls EPM (entry and product market) competition. In EPM, diminishing returns to scale need never hold as in conventional pure competition, but freedom of entry for tiny interstitial OMF is an adequate check on the abuses of monopoly and oligopoly—when coupled, I should myself add, with plausible threats of regulation or socialization. (Lydall sees this EPM-OMF framework as among his major contributions.) The EPM competition is atomistic but imperfect. It permits, in particular, imperfections in knowledge, unutilized economies of scale, and heterogeneous capital. It is consistent even in the long run, Lydall believes, with such phenomena as markup pricing, production under increasing returns, and persistent differentials in rates of return on investments of different sorts. As to the importance of these modifications, my reaction is wait and see, influenced more by the post-Chamberlin history of the "large group case" of monopolistic competition than by any analytical flaw in EPM-OMF or its applications.

Two further innovations in Lydall's approach and results relate to the dynamics of distribution and development. They explore the Kuznets conjecture that, in Lydall's words (pp. 129, 130), "the degree of inequality is fairly low in the poorest countries, increases with per capita income up to a certain level, and thereafter follows a continuous downward trend." In a cross-country study, a Lydall student (T. Das) derived an empirical estimate of the determinants of the Gini concentration ratio, G, with dependent variables X (income per capita), its reciprocal, and four D_i or "dummy" variables:[1] $G = 0.784 - 0.050 \log X - 12.184X^{-1} + 0.051D_1 + 0.124D_2 - 0.033D_3 - 0.007D_4$.

[1] I interpret D_1 as the proportion of income recipients (to total population), D_2 as the proportion of economically active persons, D_3 as the proportion in the rural (or

Lydall provides later (pp. 219, 220) his own explanation of the Kuznets phenomenon. He has divided each economy under study into a traditional sector, 1, and a modern sector, 2. The average income, Y_i, of each sector is related logarithmically to overall average income, $\bar{Y}$: $\log Y_1 = 1.457 + 0.725 \log \bar{Y}$, $\log Y_2 = 3.86 + 0.55 \log \bar{Y}$; we also have $\bar{Y} = (1 - p)Y_1 + pY_2$, with the proportion p rising as development progresses. Lydall assumes income distributed lognormally in each sector with a uniform logarithmic standard deviation. From these results and assumptions he derives by simulation (his table 12.1) numerical results which not only conform to the Kuznets hypothesis but go beyond it in one important respect. The share of the lowest 20 percent of the population eventually turns up as development proceeds, which it did not appear to do in Kuznets's original data.

The most noteworthy feature of Lydall's discussion of personal income distribution (chaps. 12–14) is its conscious endeavor to redress the balance between competitors already in the field. As economists know, the human-capital theories of T. W. Schultz, Gary Becker, and others have risen to dominance over the past quarter century. Lydall represents a reasoned and good-natured rebuttal in the ongoing debate. At the same time, he avoids what I should call the major error of Jan Tinbergen's 1975 distribution monograph, which is to ignore nonlabor incomes entirely as atavistic residuals destined to depart the scene in the near future.

Postwar distribution theory has been marked by a number of efforts to bridge the gaps and construct at least pontoon bridges between deductive functional theories and empirical personal ones. These efforts have ranged in America alone from the "human-capital" approach mentioned above to "job-ladder" and "labor-market segmentation" approaches generally more critical of the status quo. Lydall's study is a major and possibly a highly influential addition to this literature, although Lydall's own summary of his innovative conclusions (chap. 15, esp. pp. 285–91) is occasionally less than generous to his numerous predecessors. If Lydall's book is intended as more than a contribution, but as the late twentieth-century equivalent of John Bates Clark's late nineteenth-century magnum opus (as that summary chapter hints), I fear that it falls somewhat short.

Martin Bronfenbrenner

Duke University

agricultural) sector, and D_4 as the proportion in the urban (or nonagricultural) sector. (The statistical significance of D_3 and D_4 appears dubious in Lydall's text.) The R^2 value of 0.38, while low, is significantly greater than zero.